Exploiting Historical Databases to Design the Target Quality Profile for a New Product

ARTICLE *in* INDUSTRIAL & ENGINEERING CHEMISTRY RESEARCH · MAY 2013

Impact Factor: 2.59 · DOI: 10.1021/ie3032839

Pierantonio Facco
University of Padova

Salvador Garcia-Munoz
Eli Lilly

Fabrizio Bezzo
University of Padova

I&EC research
Industrial & Engineering Chemistry Research

Article

pubs.acs.org/IECR

Exploiting Historical Databases to Design the Target Quality Profile for a New Product

Emanuele Tomba,[†] Pierantonio Facco,[†] Fabrizio Bezzo,[†] and Salvador García-Muñoz[‡,*]

[†]CAPE-Lab—Computer-Aided Process Engineering Laboratory, Dipartimento di Ingegneria Industriale, Università di Padova, via Marzolo 9, 35131 Padova PD, Italy

[‡]Pfizer Worldwide R&D, 445 Eastern Point Road, Groton, Connecticut 06340, United States

ABSTRACT: Latent variable regression model (LVRM) inversion has been demonstrated to be a valid tool to support the design of new products and processes and for process control. One of the basic assumptions of LVRM inversion is that the set of desired characteristics specified for the inversion must adhere to the covariance structure of the historical data used to build the model. When developing a new product, it is not unusual that the desired product characteristics are assigned or allowed to vary slightly so as to satisfy the end-customer or the downstream processing requirements. If these values do not obey the covariance structure described by the LVRM, a mismatch between the model estimates obtained from the inversion solution and the desired product properties is observed. In this paper we address the above-mentioned issue: starting from a set product characteristics that does not comply with the LVRM structure, we propose a strategy to assist the selection of the new product quality profile, which is most suitable for LVRM inversion. Two approaches exploiting the desired values for the product characteristics and the LVRM parameters are proposed to reconstruct a new product profile in order to minimize the mismatch between model estimates and desired product properties. The feasibility of the proposed methodology is tested in a pharmaceutical product development case study, where the product is obtained through high-shear wet granulation.

1. INTRODUCTION

Stemming from the original work of Jaeckle and MacGregor,[1,2] the use of the inverse of a latent variable regression model (LVRM) has been demonstrated to be a useful tool to support decision making processes in several applications, ranging from the design of operating conditions,[1−4] to LVM-based predictive control,[5−7] product optimization[8] and product formulation design.[9] In general, the objective of LVRM inversion is to estimate a set of input variables $\mathbf{x}^{\mathrm{NEW}}$ ensuring the achievement of a desired set of output variables $\mathbf{y}^{\mathrm{DES}}$. In this paper, we consider the problem of the development of a new product, whose quality profile is characterized by a set $\mathbf{y}^{\mathrm{DES}}$ of properties, some of which may be assigned *a priori*, whereas some other may be allowed to vary slightly.

The studies on LVRM inversion[3−10] are built on an optimization framework aiming at minimizing the difference between the desired $\mathbf{y}^{\mathrm{DES}}$ and the one estimated through model inversion. This difference can either be forced through a hard equality constraint, or added as a weighted term to the objective function, which we will refer to as a "soft constraint". Any difference resulting between the optimal solution ($\hat{\mathbf{y}}^{\mathrm{NEW}}$) and the desired one (assuming no other constraint is active) will be proportional to the orthogonal distance between the assigned values of $\mathbf{y}^{\mathrm{DES}}$ and its projection onto the model hyperplane of the latent variables. Thus, to obtain a solution close to the target, the desired set of quality specifications $\mathbf{y}^{\mathrm{DES}}$ should adhere to the covariance structure of the historical data used to build the underlying model.[1] This, however, may become an issue when the end customer assigns specific values to some of the elements of $\mathbf{y}^{\mathrm{DES}}$: in this case, it is necessary to iterate between the customer requirements and a feasible $\mathbf{y}^{\mathrm{DES}}$ complying with the covariance structure of the $\mathbf{Y}$ matrix used to train the model. Clearly, there is a potential conflict between the importance assigned by the end user to the elements of $\mathbf{y}^{\mathrm{DES}}$, and the ability of the model to represent them.

If the desired values of $\mathbf{y}^{\mathrm{DES}}$ assigned by the user were set as hard rather than soft constraints (i.e., a weighted term in the objective function) in the optimization framework, then the possibility arises that the assigned values of $\mathbf{y}^{\mathrm{DES}}$ will not lie on the model hyperplane. Therefore, the optimization step may fail due to the possible infeasibility of the hard equality constraints, since it may be numerically impossible to obtain the desired values for the elements of $\mathbf{y}^{\mathrm{DES}}$ and simultaneously satisfy the covariance structure of $\mathbf{Y}$ described by the model loadings.

In this paper we address the above-mentioned challenges by proposing a structured approach to guide the reassessment of a preliminarily defined target attribute profile $\mathbf{y}^{\mathrm{DES}}$ in order to obtain a new product quality profile with the same covariance structure as the matrix of historical response variables used to build the model. Given such a vector for $\mathbf{y}^{\mathrm{DES}}$, it is then possible to use hard rather than soft constraints into the optimization formulation for the LVRM inversion, thus making the solution less computationally expensive. Two scenarios are taken into account: (i) the end-user wishes to specify *some* elements of $\mathbf{y}^{\mathrm{DES}}$ only, and (ii) the end-user wishes to specify *all* elements of $\mathbf{y}^{\mathrm{DES}}$.

The paper is organized as follows: in section 2 we briefly discuss the inversion of latent variable regression models and further elaborate on the mathematics of the problem addressed here; in section 3 we describe the proposed methods to select $\mathbf{y}^{\mathrm{DES}}$;

Received: November 28, 2012
Revised: April 2, 2013
Accepted: May 13, 2013
Published: May 13, 2013

the applicability of the approach is demonstrated in section 4 through a case study; eventually some final remarks conclude the work.

2. INVERSION OF LATENT VARIABLE REGRESSION MODELS

2.1. Latent Variable Models. Latent variable regression models (such as partial least-squares regression, PLS[11]) are data-based models that relate regressor and response matrices by projecting them onto a latent space of a much smaller rank (the model space) than the original matrices. Given a data set $\mathbf{X}$ $[I \times N]$ of I different input conditions (raw materials or process parameters) in which N variables are measured (the regressor space) and a data set $\mathbf{Y}$ $[I \times M]$ in which the M product properties for the I products are collected (the response space), after the appropriate pretreatment is applied to the data (e.g., mean-centering and scaling them), the LVRM represents the matrices in terms of their latent variables:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}_{\mathbf{X}} \tag{1}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{E}_{\mathbf{Y}} \tag{2}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} \tag{3}$$

The data in $\mathbf{X}$ are therefore projected onto the model latent space through the weight matrix $\mathbf{W}^*$ $[N \times A]$, giving the matrix $\mathbf{T}$ $[I \times A]$ of the projections (called scores) of the original variables in the latent space of the model. The projections in $\mathbf{T}$ are then related to the original matrices $\mathbf{X}$ and $\mathbf{Y}$ through the loading matrices $\mathbf{P}$ $[N \times A]$ and $\mathbf{Q}$ $[M \times A]$. $\mathbf{E}_{\mathbf{X}}$ $[I \times N]$ and $\mathbf{E}_{\mathbf{Y}}$ $[I \times M]$ are the residual matrices accounting for the model mismatch in the reconstruction of $\mathbf{X}$ and $\mathbf{Y}$. A represents the number of significant latent variables (LVs) chosen to build the model; namely, it corresponds to the dimension of the latent space. In general, the LVs can be interpreted as the main driving forces explaining the systematic variability in $\mathbf{X}$ that is related to the variability in $\mathbf{Y}$. For a detailed theoretical discussion on PLS, the user is referred to the papers from Höskuldsson[11] and Burnham et al.[12]

2.2. Latent Variable Model Inversion Procedure. The objective of model inversion is that of using a model to estimate a new regressor vector $\mathbf{x}_{\mathrm{NEW}}$ corresponding to a desired target quality attribute profile of response variables ($\mathbf{y}^{\mathrm{DES}}$). In general, a LVRM inversion exercise goes through the following steps:

1. build the LVRM between the preprocessed $\mathbf{X}$ and $\mathbf{Y}$ matrices;
2. set the possible constraints or physical bounds for $\mathbf{y}^{\mathrm{DES}}$ and $\mathbf{x}^{\mathrm{NEW}}$;
3. if multiple equality constraints are given for $\mathbf{y}^{\mathrm{DES}}$ (or for $\mathbf{x}^{\mathrm{NEW}}$), verify that these values are feasible within the model space;
4. invert LVRM solving the appropriate inversion problem.

The central exercise in the inversion step (step no. 4) is that of identifying a score vector ($\boldsymbol{\tau}$) that will result, during prediction and reconstruction of $\mathbf{x}^{\mathrm{NEW}}$ and $\hat{\mathbf{y}}^{\mathrm{NEW}}$, in the desired values for the elements of $\mathbf{y}^{\mathrm{DES}}$ as well as fulfilling any constraints identified either in the elements of the target attribute product profile ($\mathbf{y}^{\mathrm{DES}}$) or in the elements of the regressor ($\mathbf{x}^{\mathrm{NEW}}$). Mathematically, there are two different ways to reach a solution for $\boldsymbol{\tau}$ that satisfies the equality constraints in $\mathbf{x}^{\mathrm{NEW}}$ and $\mathbf{y}^{\mathrm{DES}}$: (i) to set soft constraints in the formulation of the objective function using ad-hoc weights; (ii) to set hard constraints for these equalities (see eqs 7 and 8 in Tomba et al.[10]).

Assigning soft constraints in the objective function offers the advantage that the equality constraints specified in either $\mathbf{x}^{\mathrm{NEW}}$ or $\mathbf{y}^{\mathrm{DES}}$ do not necessarily lie in the model subspace for the optimizer to find a solution. There are two main downsides in using soft constraints in the model inversion optimization problem. The first one is the inherent need for the user to define weights for each of the terms of the objective function and for each of the elements in $\mathbf{x}^{\mathrm{NEW}}$ and $\mathbf{y}^{\mathrm{DES}}$. Second, soft constraints add additional degrees of freedom to the optimizer, thus making the optimization exercise harder. Conversely, the establishment of hard equality constraints for the desired values of elements in $\mathbf{x}^{\mathrm{NEW}}$ and $\mathbf{y}^{\mathrm{DES}}$ reduces the number of iterations in the application of LVRM inversion procedure to support the design problem, since once the solution is found there is no discrepancy between the obtained and the desired values of the response $\mathbf{y}$. Moreover, the problem is easier to solve for the optimizer from a numerical point of view. The downside of using hard constraints in the LVRM inversion problem is that the set of constraints given for $\mathbf{x}^{\mathrm{NEW}}$ and $\mathbf{y}^{\mathrm{DES}}$ must be coherent with the covariance structure of the original matrices used to build the model. In this context, a methodology would be needed to allow a product developer to set as many hard constraints for the elements of the desired product profile $\mathbf{y}^{\mathrm{DES}}$ as possible, while adhering to the historical data covariance structure described by the model. Two possible solutions to address this problem are proposed in the following section.

2.3. Residual Space in a Latent Variable Model. As highlighted in step 3, if the specified values for $\mathbf{y}^{\mathrm{DES}}$ are outside the model subspace (described by the $\mathbf{Q}$ loadings matrix), the solution obtained from the inversion problem will not be able to provide the desired product quality profile. A metric that can be used to quantify the distance of $\mathbf{y}^{\mathrm{DES}}$ from the model space is the squared prediction error ($\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$) which is defined as:[13]

$$\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}} = \sum_{i=1}^{M} (y_i^{\mathrm{DES}} - \hat{y}_i^{\mathrm{DES}})^2 \tag{4}$$

where $\hat{y}_i^{\mathrm{DES}}$ is the ith element of the vector $\hat{\mathbf{y}}^{\mathrm{DES}}$, which represents the projection of $\mathbf{y}^{\mathrm{DES}}$ through the LVRM. In principle, it is desired that this distance be zero so that the hard constraint $y_i^{\mathrm{DES}} = \hat{y}_i^{\mathrm{DES}} = \mathbf{q}_i\boldsymbol{\tau}$ for the ith variable specified for $\mathbf{y}^{\mathrm{DES}}$ can be established, where $\mathbf{q}_i$ represents the ith row of the $\mathbf{Q}$ matrix.[10] Figure 1 illustrates a geometrical interpretation of the SPE metric.

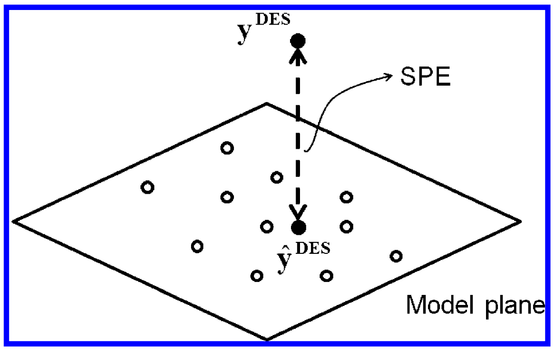

Figure 1. Representation of the geometrical interpretation of the SPE metric.

Indeed, when the target profile $\mathbf{y}^{\mathrm{DES}}$ (or a subset of it) of a new product is identified by the product development team, it may happen that this profile does not entirely match the covariance structure of the historical samples in **Y**. In this case, $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ is greater than zero and the product quality set $\hat{\mathbf{y}}^{\mathrm{NEW}}$ corresponding to the LVRM inversion solution (section 2.2) will differ from the specified $\mathbf{y}^{\mathrm{DES}}$ if a soft constraint is assigned, or in an infeasible problem if a hard constraint on $\mathbf{y}^{\mathrm{DES}}$ is assigned.[10]

For $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ to be approximately zero, $\hat{\mathbf{y}}^{\mathrm{DES}}$ should be used instead of $\mathbf{y}^{\mathrm{DES}}$ for the model inversion. Two alternatives can then be considered: (i) estimate $\hat{\mathbf{y}}^{\mathrm{DES}}$ by directly projecting and reconstructing $\mathbf{y}^{\mathrm{DES}}$ through the model; (ii) force some of the elements of $\hat{\mathbf{y}}^{\mathrm{DES}}$ to be equal to those assigned in $\mathbf{y}^{\mathrm{DES}}$. In the first case, $\hat{\mathbf{y}}^{\mathrm{DES}}$ can be very different from the desired product quality set $\mathbf{y}^{\mathrm{DES}}$, despite being its best reconstruction on the model space (minimum distance from $\mathbf{y}^{\mathrm{DES}}$). Even if this allows the use of the above-mentioned hard constraint on $\mathbf{y}^{\mathrm{DES}}$ for the model inversion, none of the product quality variables in $\hat{\mathbf{y}}^{\mathrm{DES}}$ will have the same value as that originally specified in $\mathbf{y}^{\mathrm{DES}}$. As a consequence, the product quality set $\hat{\mathbf{y}}^{\mathrm{NEW}}$ obtained from the LVRM inversion may be significantly different from $\mathbf{y}^{\mathrm{DES}}$.

In the second case, since the interest is to satisfy the constraints for the desired values of the elements of $\mathbf{y}^{\mathrm{DES}}$ as closely as possible, some of the elements of $\hat{\mathbf{y}}^{\mathrm{DES}}$ are imposed to be equal to those assigned in $\mathbf{y}^{\mathrm{DES}}$, while estimating a proper value for the others (e.g., the conditional mean[14,15]).

In the following section we will illustrate two different strategies for the selection of $\hat{\mathbf{y}}^{\mathrm{DES}}$ based on the second alternative described above. The strategies differ according to the way in which the specifications for the elements in $\mathbf{y}^{\mathrm{DES}}$ (namely, the equality constraints) are managed. In the first approach, one of the elements of $\mathbf{y}^{\mathrm{DES}}$ is assigned at a time, while the other elements are calculated through the model using a direct model inversion approach in order to obtain $\hat{\mathbf{y}}^{\mathrm{DES}}$ belonging to the model space. In the second approach, $\hat{\mathbf{y}}^{\mathrm{DES}}$ is calculated by assigning the largest number of elements of $\mathbf{y}^{\mathrm{DES}}$ still leading to obtain a $\hat{\mathbf{y}}^{\mathrm{DES}}$ within the model space; the number and type of the elements are selected according to an optimal criterion. In both approaches it is assumed that the values for all the M elements of $\mathbf{y}^{\mathrm{DES}}$ have been assigned by the user, but the methods can be easily applied even if $S < M$ elements have been specified. Additional details on the models to use for the reconstruction of $\hat{\mathbf{y}}^{\mathrm{DES}}$ and on the difference in reconstructing $\hat{\mathbf{y}}^{\mathrm{DES}}$ with a PCA or a PLS model are given in Appendix A. It is proper to mention that all the discussion and methodology presented in this work can also be applied to select equality constraints for $\mathbf{x}^{\mathrm{NEW}}$, since it is important for any fixed elements in this vector to exhibit the same covariance structure as in the **X** matrix used to build the model.

3. USING THE HISTORICAL KNOWLEDGE TO SELECT THE PRODUCT TARGET ATTRIBUTE PROFILE

3.1. Theoretical Considerations. Let us consider the vector of the desired product profile $\mathbf{y}^{\mathrm{DES}}$ autoscaled according to the mean and the standard deviation of the columns of the historical dataset of product properties (**Y**) used to build the model. Although the method outlined subsequently in this section (Method 1) is proposed to provide an estimate of the $M - 1$ free elements of $\mathbf{y}^{\mathrm{DES}}$, given an equality constraint enforced on each ith element y_i^{DES} of $\mathbf{y}^{\mathrm{DES}}$, it is reasonable to think that such an estimate is uniquely defined only if the $M - 1$ free elements of $\mathbf{y}^{\mathrm{DES}}$ have a strong correlation with the ith that is being assigned. This situation would imply that the effective rank of **Y** is 1 (only one LV is necessary to represent **Y**) and hence the reconstruction of $\hat{\mathbf{y}}^{\mathrm{DES}}$ based on one element is reasonable.

However, if the effective rank of **Y** is of higher order, assigning the value of the ith element of $\mathbf{y}^{\mathrm{DES}}$ would create an *induced null space* where multiple values of $\boldsymbol{\tau}$ can provide the same predicted value for y_i^{DES} while providing multiple possible values for the $M - 1$ free elements of $\mathbf{y}^{\mathrm{DES}}$, depending on the correlation structure of **Y**. This artificial null space will change depending on what element of $\mathbf{y}^{\mathrm{DES}}$ is being assigned and can be explicitly determined by the linear system of equations $\hat{\mathbf{y}}^{\mathrm{DES}} = \mathbf{Q}\boldsymbol{\tau}$ represented by:

$$\begin{array}{lllllll}
\hat{y}_1^{\mathrm{DES}} = & q_1^1\tau_1 & +q_1^2\tau_2 & +q_1^3\tau_3 & \cdots & +q_1^A\tau_A & \\
\hat{y}_2^{\mathrm{DES}} = & q_2^1\tau_1 & +q_2^2\tau_2 & +q_2^3\tau_3 & \cdots & +q_2^A\tau_A & \\
\hat{y}_3^{\mathrm{DES}} = & q_3^1\tau_1 & +q_3^2\tau_2 & +q_3^3\tau_3 & \cdots & +q_3^A\tau_A & , \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \\
\hat{y}_m^{\mathrm{DES}} = & q_m^1\tau_1 & q_m^2\tau_2 & q_m^3\tau_3 & \cdots & q_m^A\tau_A &
\end{array} \tag{5}$$

q_i^j in (5) being the element on the ith row and jth column of the **Q** matrix, and τ_j the jth element of the column vector $\boldsymbol{\tau}$, with $i = 1, ..., M$ and $j = 1, ..., A$. For illustrative purposes, consider four scenarios:

(a) the variables of **Y** are completely independent (**Y** is full rank) and each variable is represented by one LV;

(b) **Y** is full rank, but the variables are explained in groups by different components;

(c) **Y** is rank deficient and correlation is captured in A components, where $A < M$;

(d) **Y** is rank deficient and correlation is captured in A components, where $A \le M$, but $A > \mathrm{rank}(\mathbf{Y})$.

The appearance of these induced null spaces is obvious in case (a), where the right-hand of eq 5 is reduced to the main diagonal elements (eq 6). If an equality constraint is applied to one of the elements of $\mathbf{y}^{\mathrm{DES}}$, the rest of the equations define the available null space in which the user can pick any value of the free scores.

$$\begin{array}{lllll}
\hat{y}_1^{\mathrm{DES}} = & q_1^1\tau_1 & & & \\
\hat{y}_2^{\mathrm{DES}} = & & q_2^2\tau_2 & & \\
\hat{y}_3^{\mathrm{DES}} = & & & q_3^3\tau_3 & \\
\vdots & & & & \ddots \\
\hat{y}_m^{\mathrm{DES}} = & & & & q_m^A\tau_A
\end{array} \tag{6}$$

Scenario (b) would imply that there are as many independent directions of variability in **Y** as columns in it; however, each direction of change affects multiple variables. Consider the scenario of a three-dimensional space with three LVs such that the representative loadings across all LVs are as in eq 7. Given the below situation and a constraint placed on the first element of $\mathbf{y}^{\mathrm{DES}}$, one could isolate τ_1 from the first row and replace it into the third row to end with a system of two equations with four unknowns (τ_2,τ_3,$\hat{y}_2^{\mathrm{DES}}$ and $\hat{y}_3^{\mathrm{DES}}$). This system represents the two-dimensional induced null space where any solution chosen for τ_2 or τ_3 can be used to estimate τ_1 and satisfy the equality condition for the first element of $\mathbf{y}^{\mathrm{DES}}$ while also keeping the vector $\hat{\mathbf{y}}^{\mathrm{DES}}$ in the latent space.

$$\begin{array}{llll}
\hat{y}_1^{\mathrm{DES}} = & q_1^1\tau_1 & & + q_1^3\tau_3 \\
\hat{y}_2^{\mathrm{DES}} = & & q_2^2\tau_2 & \\
\hat{y}_3^{\mathrm{DES}} = & q_3^1\tau_1 & & + q_3^3\tau_3
\end{array} \tag{7}$$

In scenario (c), the number of LVs is lower than the number of variables in **Y** and not all variables are represented in all latent spaces (e.g., equation system 8). Given the situation below, a hard constraint enforced on the first element will assign the value of the fourth and will result in a one-dimensional induced null space where the user can choose any value of τ_1 (which in turn defines the values of the second and third element of $\hat{\mathbf{y}}^{\mathrm{DES}}$).

$$\begin{array}{lll}
\hat{y}_1^{\mathrm{DES}} = & & q_1^2\tau_2 \\
\hat{y}_2^{\mathrm{DES}} = & q_2^1\tau_1 & \\
\hat{y}_3^{\mathrm{DES}} = & q_3^1\tau_1 & \\
\hat{y}_4^{\mathrm{DES}} = & & q_4^2\tau_2
\end{array} \tag{8}$$

Finally, consider scenario (d). In this case the number of LVs is lower than the number of variables in $\mathbf{Y}$, however it is greater than the effective rank of $\mathbf{Y}$. This is a typical situation when building PLS models between a regressor data set $\mathbf{X}$ and a response data set $\mathbf{Y}$, and the rank of the $\mathbf{X}$ matrix is larger than the rank of the $\mathbf{Y}$ matrix. In these cases, if the PLS model is built with $A = \mathrm{rank}(\mathbf{X})$ LVs, there will be $k = A - \mathrm{rank}(\mathbf{Y})$ directions in the latent space which have little (theoretically none) influence on the $\mathbf{Y}$ space, but that are needed to adequately describe the variability in the regressor space. These latent directions form the PLS null space[1] which gives additional degrees of freedom in the estimation of the score vector $\boldsymbol{\tau}$, in addition to the induced null spaces which can be generated in situations similar to those described above. For example, assume the same case as in eq 8 in which the $\mathbf{Y}$ space is four-dimensional and $\mathrm{rank}(\mathbf{Y}) = 2$, but consider that three LVs were chosen to build the PLS model, as they were needed to represent adequately the $\mathbf{X}$ space (eq 9). In this case, a hard constraint imposed on the first element will result in the estimation of τ_2, but the user can choose any value for τ_1 and τ_3. The system represents a two-dimensional null space, which however is formed by the combination of a one-dimensional induced null space and the PLS null space due to the differences in the ranks of $\mathbf{X}$ and $\mathbf{Y}$.

$$\begin{array}{lllll}
\hat{y}_1^{\mathrm{DES}} = & & q_1^2\tau_2 & & \\
\hat{y}_2^{\mathrm{DES}} = & q_2^1\tau_1 & & & \\
\hat{y}_3^{\mathrm{DES}} = & q_3^1\tau_1 & & + & q_3^3\tau_3 \\
\hat{y}_4^{\mathrm{DES}} = & & q_4^2\tau_2 & + & q_4^3\tau_3
\end{array} \tag{9}$$

Note that the case presented in scenario (d) (eq 9) could only occur when $\hat{\mathbf{y}}^{\mathrm{DES}}$ is reconstructed through the PLS $\mathbf{Q}$ loadings, differently from the situations described in the previous scenarios which are valid also in the cases in which $\hat{\mathbf{y}}^{\mathrm{DES}}$ is reconstructed based on the PCA loadings (Appendix A).

In practice, it is common to have only desirable ranges for some quality attributes of the product while having specific assigned conditions for other quality descriptors. In this paper we suggest to handle the free elements of $\mathbf{y}^{\mathrm{DES}}$ as missing data for the sake of simplicity and to expedite the decision of a vector of quality properties that will result in the target overall performance for a new product. Other researchers have already presented and studied the behavior of the analytical estimators for each of the missing data methods and have already discussed whether the expected predicted value for the missing elements are close to the unconditional mean, the conditional mean, the least Mahalanobis distance, or the least squared prediction error.[15,14] From the perspective of this application, we use similar approaches as a shortcut to the construction of potential vectors representing the target quality profile for a product. The discussion on the induced null space is presented due to the obvious reaction toward predicting the majority of the $\hat{\mathbf{y}}^{\mathrm{DES}}$ vectors based on one element (in the worst of the cases, these predictions will be nothing but the unconditional means).

In the following sections the algorithms on which the proposed methods are based are described. The first method (Method 1) offers a simple procedure to appreciate the trade-offs and understand the effects that assigning one element of the quality profile versus another one has on the product quality profile. Differently, the second method (Method 2) provides a procedure to find an optimal product quality profile as a trade-off between the historical data covariance structure and the user-defined values of the quality variables.

3.2. Method 1: Assigning One Quality Variable at a Time. In the first method, the selected product quality set $\hat{\mathbf{y}}^{\mathrm{DES}}$ is calculated imposing that, for the ith element of $\hat{y}^{\mathrm{DES}}$, $\hat{y}_i^{\mathrm{DES}} = y_i^{\mathrm{DES}}$, while the values of the other elements in $\hat{\mathbf{y}}^{\mathrm{DES}}$ are assumed to be missing.

Several approaches have been proposed to deal with missing data when using multivariate statistical techniques like PCA or PLS.[14,15] In all these contributions the objective was to use the model to estimate the scores corresponding to a new sample presented to the model characterized by missing measurements. In general this is concerned with multivariate statistical process control or process modeling applications, where missing measurements are in the input data (namely in the regressor side, if a PLS model is considered). Differently, in this application, in order to reconstruct the product quality profile we are considering as missing the measurements referred to the response set $\hat{\mathbf{y}}^{\mathrm{DES}}$, thus using the model inversion to reconstruct them on the basis of the available fixed values. The proposed method exploits the submodel constituted by the PLS $\mathbf{Q}$ loadings to reconstruct the new product target profile $\hat{\mathbf{y}}^{\mathrm{DES}}$ through a direct inversion of the PLS model.[1]

For each variable i, the proposed procedure aims at estimating the scores $\hat{\boldsymbol{\tau}}^{(i)}$ of $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ on the basis of the ith element (y_i^{DES}), which is assigned, by projecting it back to the model plane through a direct inversion of the model:

$$\boldsymbol{\tau}^{(i)} = (\mathbf{Q}^{(i)\mathrm{T}}\mathbf{Q}^{(i)})^{-1}\mathbf{Q}^{(i)\mathrm{T}}y_i^{\mathrm{DES}} \tag{10}$$

where $\mathbf{Q}^{(i)}$ is the submatrix of the loadings $\mathbf{Q}$ in which only the row of $\mathbf{Q}$ corresponding to the element assigned in $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ is considered. Namely, according to the procedure in Method 1, $\mathbf{Q}^{(i)}$ is a row vector of $[1 \times A]$ dimensions.

This method is applied to all the M variables specified for $\mathbf{y}^{\mathrm{DES}}$, giving then in output a matrix $\hat{\mathbf{Y}}^{\mathrm{DES}^{\mathrm{T}}}$, whose columns are the M different reconstructions for the product quality $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ obtained assigning in turn each element i. The procedure goes through the following steps:

1. Assign the value of the ith element of $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ in order for it to be equal to the corresponding element in $\mathbf{y}^{\mathrm{DES}}$ ($\hat{y}_i^{\mathrm{DES}(i)} = y_i^{\mathrm{DES}}$), considering the other elements in $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ as missing data.
2. Estimate the score vector $\hat{\boldsymbol{\tau}}^{(i)}$ corresponding to $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ through the direct model inversion in eq 10.
3. Reconstruct $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ from $\hat{\boldsymbol{\tau}}^{(i)}$ and the $\mathbf{Q}$ loadings of the PLS model and store it in the matrix $\hat{\mathbf{Y}}^{\mathrm{DES}^{\mathrm{T}}}$:

$$\hat{\mathbf{y}}^{\mathrm{DES}(i)} = \mathbf{Q}\hat{\boldsymbol{\tau}}^{(i)} \tag{11}$$

4. Assign the next desired product property, until all the M properties in $\mathbf{y}^{\mathrm{DES}}$ have been considered.

Thus, from the different suggestions for the new product quality set $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$, the user can have an idea of the general impact that a change in each of the original variables will have to a change in the covariance of the desired quality profile so that it better resembles that of the historical data. The user would then select the candidate solution that better adheres to the end-customer needs.

In some cases, reconstructing $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ through the direct inversion of the model (eq 10) may lead to an unfeasible solution clashing against the physical limits some of the product quality variables may have. If that occurs, additional flexibility to the procedure described for Method 1 may be added by substituting the step 2 and 3 of the above procedure (i.e., eq 10 and eq 11) with the optimization problem described in the next section (solving eq 13 instead). That will be better clarified when discussing the case study results in section 4.2.

3.3. Method 2: Assigning More than One Quality Variable.

The second proposed method is based on an approach that is somehow dual to Method 1. The method starts from the originally defined product quality vector $\mathbf{y}^{\mathrm{DES}}$, and uses an iterative procedure to progressively find the assigned variables in $\mathbf{y}^{\mathrm{DES}}$ that contribute the most to the $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ value, and to remove the corresponding equality constraints until a new estimated desired product quality $\mathbf{y}^{\mathrm{NEW}}$ is obtained that is as close as possible to the model space. Let us assume that the variables in $\mathbf{y}^{\mathrm{DES}}$ are completely (or for the most part) assigned (all equality constraints) and let us define a (small) threshold ε setting the acceptability limit for the value of $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$. To ensure that the new product quality profile in $\mathbf{y}^{\mathrm{NEW}}$ belongs to the model space, the value defined for ε should be significantly lower than the SPE values of the historical data used to build the model. A general schematic of the procedure is reported in Figure 2.

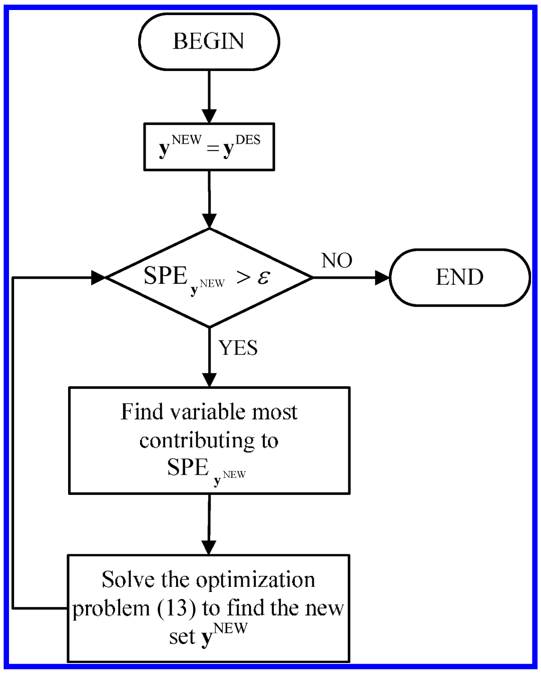
BEGIN
$\mathbf{y}^{\mathrm{NEW}} = \mathbf{y}^{\mathrm{DES}}$
$\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}} > \varepsilon$
NO
END
YES
Find variable most contributing to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$
Solve the optimization problem (13) to find the new set $\mathbf{y}^{\mathrm{NEW}}$

Figure 2. Schematic of the algorithm implemented for Method 2.

After setting $\mathbf{y}^{\mathrm{NEW}} = \mathbf{y}^{\mathrm{DES}}$, at each iteration the procedure verifies if the desired set $\mathbf{y}^{\mathrm{NEW}}$ belongs to the model space, by calculating $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ and comparing it to ε. If this is not verified, the contributions to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ are calculated according to

$$\mathrm{SPE}_{\mathrm{CONT},m} = f_m \cdot f_m \tag{12}$$

where f_m is the residual for the mth element of $\mathbf{y}^{\mathrm{NEW}}$.

The element in $\mathbf{y}^{\mathrm{NEW}}$ with the highest contribution to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ is selected and its constraint is removed. An optimization problem is then solved to update the design set $\mathbf{y}^{\mathrm{NEW}}$:

$$\begin{aligned} &\min_{\hat{\boldsymbol{\tau}}} (\mathbf{y}^{\mathrm{NEW}} - \mathbf{Q}\hat{\boldsymbol{\tau}})^{\mathrm{T}} \Gamma (\mathbf{y}^{\mathrm{NEW}} - \mathbf{Q}\hat{\boldsymbol{\tau}}) + g \cdot \sum_{a=1}^{A} \frac{\hat{\tau}_a^2}{s_a^2} \\ &\text{s.t.} \\ &y_m^{\mathrm{NEW}} = y_m^{\mathrm{DES}} \\ &y_j^{\mathrm{NEW}} < b_j \\ &lb_k < y_k^{\mathrm{NEW}} < ub_k \end{aligned} \tag{13}$$

where b_j represents the value of the inequality constraint assigned to the jth element of $\mathbf{y}^{\mathrm{NEW}}$ to account for the specification limit possibly specified for it, while lb_k and ub_k are respectively the lower and the upper physical boundaries for the k-th element of $\mathbf{y}^{\mathrm{NEW}}$. Note that physical boundaries represent the variable domain in the optimization procedure. Conversely, inequality constraints represent the regions inside which the product properties are desired to fall, and are then subsets of the physical bounds. In the objective function, $\boldsymbol{\Gamma}$ is a diagonal matrix of weights defined by the user according to the relative importance he/she wants to give to the product quality variables (usually a good choice for the elements of $\boldsymbol{\Gamma}$ is represented by the explained variance per variable ($R_{y,\mathrm{pv}}{}^2$), calculated on the historical samples); g is a weight for the second term of the objective function where $\hat{\tau}_a$ is the ath element of $\hat{\boldsymbol{\tau}}$ and s_a^2 is the variance of the ath column of matrix $\mathbf{T}$.

From the optimization problem, a new set $\mathbf{y}^{\mathrm{NEW}}$ representing the new estimated target quality profile is obtained, which is again assessed against threshold ε. The procedure of progressively removing the equality constraints initially specified for the elements of $\mathbf{y}^{\mathrm{DES}}$ is repeated until $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ is found below the given threshold ε. Then, a new product quality set $\mathbf{y}^{\mathrm{NEW}}$ $(= \hat{\mathbf{y}}^{\mathrm{DES}})$ is obtained, which represents the best compromise between the set of the target quality profile initially defined by the user ($\mathbf{y}^{\mathrm{DES}}$) and the model requirements.

Conversely, if after removing all the equality constraints in $\mathbf{y}^{\mathrm{DES}}$, $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}} > \varepsilon$ still holds, then the problem is unfeasible and a revision on the constraints specified in eq 13 should be considered.

The second term of the objective function in eq 13 represents the Hotelling's T^2 of the solution. This term is added to the objective function to consider the cases in which an induced null space is present due to the structure of the loadings $\mathbf{Q}$. In these cases, the null space can be exploited to move the solution along it, in order to find a new set $\mathbf{y}^{\mathrm{NEW}}$ that belongs to the model space, but at the same time is inside (or close to) the range of the properties of the historical products (thus avoiding extrapolated solutions). To this end, we have $g \neq 0$ and reliably $g \ll 1$ in eq 13, in order to give more importance in the objective function to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}}$ rather than to the Hotelling's T^2 of the solution. In the case the null space is due to the differences in the ranks of the $\mathbf{X}$ and $\mathbf{Y}$ matrices (namely, the $\mathbf{Q}$ loading matrix is redundant), g can be set to zero, since the Hotelling's T^2 of the solution is considered in the subsequent PLS inversion problem, for the estimation of the regressors which provide the desired responses $\mathbf{y}^{\mathrm{NEW}}$.[10] Note also that the problem in eq 13 can be solved by introducing slack variables to soften the hard equality constraints assigned to the elements of $\mathbf{y}^{\mathrm{NEW}}$.[16]

Finally, note that the solution of Method 2 coincides with that of Method 1 (namely, the direct inversion of the model) if an equality constraint for only one element of $\mathbf{y}^{\mathrm{DES}}$ is assigned in eq 13, $g = 0$ and no inequality constraints or boundaries are present in the inversion problem.

4. CASE STUDY: DEFINING THE TARGET PROFILE FOR A WET GRANULATED PRODUCT

The proposed methodologies are applied to a particle engineering problem to design the quality profile of a wet-granulated product. The ultimate objective is that of using the historical knowledge available for different products to estimate the best input raw material properties to obtain granules with the desired characteristics in output from a wet granulation process. Before performing the LVRM inversion, it must be ensured that the model is able to adequately describe the desired product quality $\mathbf{y}^{\mathrm{DES}}$, which is assigned by the user (all equality constraints on the elements of $\mathbf{y}^{\mathrm{DES}}$).

4.1. Preliminary Data Analysis. In this example the data reported in the work of Vemavarapu et al.[17] have been used. In the original work, the authors studied the influence of the raw material properties on the performances of a product under the assumption that it is obtained using a high-shear wet granulation process. Twenty-four different active pharmaceutical ingredients (APIs) were considered and processed under the same granulation conditions to study the influence of the raw material properties only. Each API was characterized by measuring seven physical properties, while the wet-granulated products were characterized by measuring seven different product characteristics. Table 1 reports the input material physical properties and the variables measured for the characterization of the products in each experimental run. Data are collected in two data sets: a data set $\mathbf{X}$ [24 × 7] including the properties of 24 APIs tested and a data set $\mathbf{Y}$ [24 × 7] of the properties of the corresponding products.

Table 1. Measured Properties for the APIs (X) and for the Wet-Granulated Products (Y)

	API properties		product properties
1.	H_2O solubility (mg/mL)	1.	loss on drying (LOD) (%)
2.	contact angle (deg)	2.	oversize (%)
3.	H_2O holding capacity (wt % gain)	3.	ΔFlodex (mm)
4.	D[3,2] (μm)	4.	ΔCompactibility (KPa/MPa)
5.	$D90/D10$	5.	D[3,2] (μm)
6.	surface area (m^2/g)	6.	$D90/D10$
7.	pore volume (cm^3/g)	7.	growth ratio

Additional details on the variables and on the rationale to select them are reported in the original paper[17] together with information on the measurement procedures and on the process conditions.

For the purpose of this study, we assume that it is required to manufacture a granulated product with the characteristics $\mathbf{y}^{\mathrm{DES}}$ reported in Table 2. Note that these data do not correspond to a real product, but in general they represent an example of the combination of desirable properties for a wet granule.

In Table 3 the diagnostics for the PLS model between the data sets $\mathbf{X}$ and $\mathbf{Y}$ are reported in terms of explained variance and cumulative explained variance per LV both in model design ($R^2\mathbf{X}$, $R^2\mathbf{Y}$, $R^2\mathbf{X}_{\mathrm{CUM}}$, $R^2\mathbf{Y}_{\mathrm{CUM}}$) and in cross-validation (P^2, Q^2, P^2_{CUM}, Q^2_{CUM}) for all the 7 LVs which can be estimated from the analysis. Note that the cross-validation diagnostics are reported not only for the response variables in $\mathbf{Y}$ (Q^2, Q^2_{CUM}), but also for the regressors in $\mathbf{X}$ (P^2, P^2_{CUM}).[10] This is due to the fact that since the objective of model inversion is the estimation of the regressors starting from the response variables, the model must ensure an adequate representation not only of the $\mathbf{Y}$, but also of the $\mathbf{X}$ space. For this reason, a comprehensive strategy for the selection of the number of LVs to design a LVRM for inversion should also consider a metric to diagnose the performances of the model in representing the $\mathbf{X}$ data.[10]

From Table 3, both $R^2\mathbf{X}$ and P^2 show a significant decrease in the amount of explained variance after the fourth LV. From the values of $R^2\mathbf{Y}$ and especially of Q^2 it can be seen that the amount of variance for $\mathbf{Y}$ explained by the LVs after the third does not seem to be significant. By considering the diagnostics for $\mathbf{Y}$ three LVs would then be sufficient to build the LVRM. However, $R^2\mathbf{X}$ and P^2 values show that the latent space for $\mathbf{X}$ is four dimensional. Accordingly, four LVs were selected for the model design. This means that there is a one-dimensional null space in the latent space of the model, which has to be considered in the inversion of the PLS model. This case study can thus potentially represent an example of Scenario (d) described in Section 3.1, as the effective rank of the $\mathbf{Y}$ space is lower than both the dimension of $\mathbf{Y}$ and the number of considered LVs. The presence of the PLS null space and of the possible induced null spaces generated in the application of the proposed procedures should then be considered in the reconstruction of $\hat{\mathbf{y}}^{\mathrm{DES}}$.

In Figure 3 the loadings $\mathbf{q}$ (i.e., the columns of the $\mathbf{Q}$ matrix in eq 2) of the PLS model on the four considered LVs are reported. The loadings have been standardized by $R^2_{y,\mathrm{pv}}$ to get a better contrast and to allow for a "cross-component" analysis. The analysis of these plots allows assessing the main driving forces, which explain the variability in the $\mathbf{Y}$ data most related to $\mathbf{X}$. These driving forces will be exploited by the proposed methodologies for the selection of the new target product quality profiles $\hat{\mathbf{y}}^{\mathrm{DES}}$. For the sake of conciseness, a detailed description and interpretation of the loading plots of Figure 3 is reported in Appendix B.

Figure 4 reports the values of the SPE versus the values of the Hotelling's T^2 for the historical samples in $\mathbf{Y}$ (black dots), together with the relevant 95% (red short-dashed lines) and 99% (blue dashed lines) confidence limits. The squared dot (□) represents the values of $T^2_{\mathbf{y}^{\mathrm{DES}}}$ and $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$. As can be seen, even if $T^2_{\mathbf{y}^{\mathrm{DES}}}$ is inside the 95% (= 10.97) confidence limit, meaning that the values specified in $\mathbf{y}^{\mathrm{DES}}$ are not far from the mean of the historical values, $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$ is above the 95% relevant limit (= 5.75), meaning that the historical correlation structure is not valid for $\mathbf{y}^{\mathrm{DES}}$. Thus, the model is not really appropriate in representing $\mathbf{y}^{\mathrm{DES}}$ due to the high model mismatch, and it is not recommended to perform the inversion with $\mathbf{y}^{\mathrm{DES}}$.

The methods previously described can then be applied to exploit the historical covariance structure of the data in order to give suggestions on possible new product properties sets $\hat{\mathbf{y}}^{\mathrm{DES}}$, which can be feasibly used for the model inversion.

Table 2. Desired Product Properties $\mathbf{y}^{\mathrm{DES}}$ for a Wet-Granulated Product

	LOD (%)	oversize (%)	ΔFlodex (mm)	ΔCompactibility (KPa/MPa)	D[3,2] (μm)	$D90/D10$	growth ratio
$\mathbf{y}^{\mathrm{DES}}$	3	0	20	5	400	2.5	8

Table 3. Diagnostics of the PLS Model between X and Y

LV	R^2X	R^2X$_{\mathrm{CUM}}$	P^2	P^2_{CUM}	R^2Y	R^2Y$_{\mathrm{CUM}}$	Q^2	Q^2_{CUM}
1	40.84	40.84	35.71	35.71	32.34	32.35	24.68	24.68
2	18.69	59.54	18.26	53.97	23.34	55.69	24.31	48.99
3	15.99	75.53	15.38	69.35	8.30	63.99	11.72	60.71
4	17.45	92.98	22.03	91.38	1.58	65.57	2.30	63.02
5	2.49	95.47	3.66	95.04	2.77	68.34	2.88	65.90

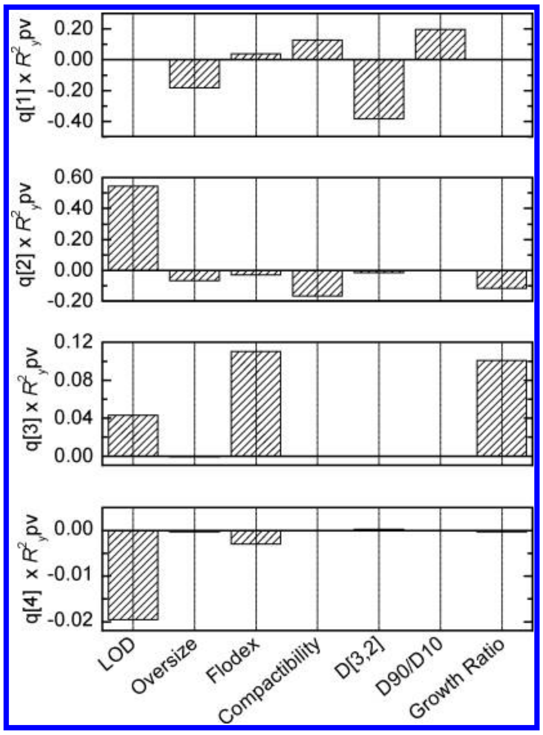

Figure 3. Loadings **Q** of the PLS model on the **X** and **Y** data sets.

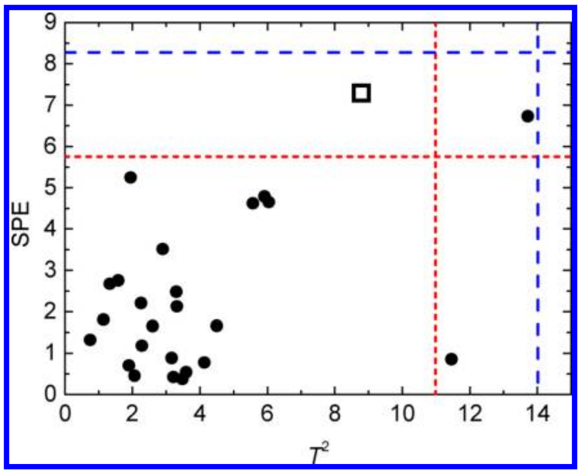

Figure 4. Plot of SPE versus T^2 values for the historical products in **Y** (●) and the new desired product quality set $\mathbf{y}^{\mathrm{DES}}$ (□). The lines represent respectively the 95% (short-dashed red) and 99% (dashed blue) confidence limits.

4.2. Method 1: Results. In Table 4 the results obtained after applying Method 1 using the proposed direct inversion approach are reported. Each row represents a vector $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ suggested by the model, obtained fixing one at a time each one of the seven properties of $\mathbf{y}^{\mathrm{DES}}$ in Table 2. The assigned properties for each of the seven $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ sets are those bold in brackets in Table 4, and coincide with the original values of $\mathbf{y}^{\mathrm{DES}}$ in Table 2. For the sake of comparison, the last row reports the set $\hat{\mathbf{y}}^{\mathrm{DES}}$ obtained by directly projecting $\mathbf{y}^{\mathrm{DES}}$ onto the model space. In the two last columns of the table, the values of the T^2 and of the SPE are also reported for each calculated set.

As it can be seen, for all the suggested new product quality sets $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$, the calculated values for the assigned elements of $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$ are equal to the corresponding equality constraint $\mathbf{y}^{\mathrm{DES}}$ (reported between parentheses). Moreover, from the values of the SPE, the suggested variable combinations result to be all lying onto the model space. It is interesting to note that for each of the calculated sets $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$, the covariance structure of the historical samples in **Y** (which is basically represented by the plots in Figure 3) is optimally used in the proposed approach to estimate the new sets $\hat{\mathbf{y}}^{\mathrm{DES}(i)}$, even if only one of the seven variables is constrained in each case, as will be discussed below. Prior to such discussion, the reader is referred to Table 5, where the mean and standard deviation values of the variables in **Y** are provided for a clear comparison with the results of Table 4.

Let us consider the case in which the percentage of oversize granules was assigned to 0 ($\hat{\mathbf{y}}^{\mathrm{DES}(2)}$), which is quite different from the historical mean in Table 5. From Figure 3, it can be seen that this variable is related to $D[3,2]$ and inversely related to $D90/D10$ on LV1; furthermore, it scarcely affects the other LVs. In fact, given that in $\hat{\mathbf{y}}^{\mathrm{DES}(2)}$ the oversize percentage is lower than the historical mean, the suggested set is characterized by a low value of $D[3,2]$ and a high value of $D90/D10$ (even if within one standard deviation as from the values in Table 5). The other variables are more similar to their unconditional mean (Table 5), because they do not show a strong relation with the oversize percentage (Figure 3).

This kind of analysis can be repeated for the other sets in Table 4. In particular, in the case of $\hat{\mathbf{y}}^{\mathrm{DES}(4)}$, a value of ΔCompactibility out of the range of the historical product data was assigned. In this case, it can be observed that in order to obtain such a value of ΔCompactibility, the product needs to exhibit a low and broad PSD, and higher-than-mean oversize percentage and ΔFlodex. Furthermore, a negative (and physically meaningless) value of loss on drying (LOD) is obtained. This can be explained because Method 1 does not allow for the inclusion of physical boundaries for the variables and, as a consequence, unfeasible design outputs may be achieved sometimes (especially when extreme values are desired for some other variables). To account for this issue, as anticipated in the end of section 3.2, the problem has been solved using the procedure described for Method 1 by substituting the direct model inversion to reconstruct $\hat{\mathbf{y}}^{\mathrm{DES}(4)}$ (eq 10 and eq 11) with the optimization framework formalized in eq 13, which allows binding the variables to physically sound values. In particular, the assigned value for ΔCompactibility (ΔCompactibility = 5) has been set as a hard constraint for $\mathbf{y}^{\mathrm{NEW}}$ in eq 13.

Using an optimization framework in this case has also another advantage. By analyzing the loading plots in Figure 3, it can be seen that ΔCompactibility is mainly described by the first two LVs (Figure 3), while it does not contribute significantly on the

Table 4. Method 1. New sets of product properties $\hat{\mathbf{y}}^{\mathrm{DES(i)}}$ for a wet-granulated product suggested on the basis of the historical data model[a]

	LOD (%)	oversize (%)	ΔFlodex (mm)	ΔCompactibility (KPa/MPa)	$D[3,2]$ (μm)	$D90/D10$	growth ratio	T^2	SPE
$\hat{\mathbf{y}}^{\mathrm{DES(1)}}$	**3 (3)**	13.5	10.2	−0.58	137	8.7	12.2	0.90	∼ 0
$\hat{\mathbf{y}}^{\mathrm{DES(2)}}$	2.6	**0 (0)**	11.2	0.22	68.4	15.0	11.7	1.00	0
$\hat{\mathbf{y}}^{\mathrm{DES(3)}}$	1.4	22.9	**20 (20)**	1.47	147	12.3	30.4	1.33	0
$\hat{\mathbf{y}}^{\mathrm{DES(4)}}$	−2.1	34.7	21.6	**5.00 (5)**	111	22.6	35.4	7.64	∼ 0
$\hat{\mathbf{y}}^{\mathrm{DES(5)}}$	1.1	41.1	9.4	0.28	**400 (400)**	4.5	17.5	0.44	∼ 0
$\hat{\mathbf{y}}^{\mathrm{DES(6)}}$	1.8	38.2	8.4	−0.49	411	**2.5 (2.5)**	15.1	0.54	∼ 0
$\hat{\mathbf{y}}^{\mathrm{DES(7)}}$	1.8	21.1	6.3	0.03	166	9.6	**8.0 (8)**	0.35	∼ 0
$\hat{\mathbf{y}}^{\mathrm{DES}}$	2.3	33.3	16.7	1.94	90.4	15.7	19.8	4.01	0

[a] The bold values represent the assigned element in each set (the original values of $\mathbf{y}^{\mathrm{DES}}$ are reported in brackets). The proposed approach has been used to calculate the others. T^2 and SPE statistics are also given.

Table 5. Mean and standard deviation (st. dev.) values for the properties of the wet granulated products in the historical database Y

	LOD (%)	oversize (%)	ΔFlodex (mm)	ΔCompactibility (KPa/MPa)	$D[3,2]$ (μm)	$D90/D10$	growth ratio
mean	1.55	24.5	10.6	0.5	181	9.8	15.8
st. dev.	1.41	29.1	9.6	1.9	410	13.1	15.1

Table 6. Method 1. New sets of product properties calculated using the optimization framework in eq 13 instead of the direct model inversion in the case ΔCompactibility of the granulated product is assigned ($\hat{\mathbf{y}}^{\mathrm{DES(4)}}$), considering or not the soft constraints (SC) on T^2.[a]

	LOD (%)	oversize (%)	ΔFlodex (mm)	ΔCompactibility (KPa/MPa)	$D[3,2]$ (μm)	$D90/D10$	growth ratio	T^2	SPE
No SC on T^2	0	0	12.4	**5 (5)**	11	36.7	7.1	11.8	0
SC on T^2	0	0	28.5	**5 (5)**	22	31.5	37.5	8.4	1.8×10^{-6}

[a] The bold values in brackets represent the values assigned to ΔCompactibility. T^2 and SPE statistics are also given.

Table 7. Method 2. New set of product properties for a wet-granulated product suggested on the basis of the historical data model.[a]

	LOD (%)	oversize (%)	ΔFlodex (mm)	ΔCompactibility (KPa/MPa)	$D[3,2]$ (μm)	$D90/D10$	growth ratio	T^2	SPE
No SC on T^2	$3^{\#}$	$0^{\#}$	$20^{\#}$	5.4	3	42.7	$8^{\#}$	49.9	0
SC on T^2	0.1	16.1	17.6	3	75	20.5	25.1	2.1	0

[a] The values with the # superscript represent the variable values equal to the ones in $\mathbf{y}^{\mathrm{DES}}$. T^2 and SPE statistics are also given. SC stands for "soft constraint".

other LVs. This means that by assigning this property, it is possible to isolate the score on LV1 or on LV2, and to replace it for the reconstruction of the other variables (see eq 7 above). The scores on the other LVs can be selected independently, thus generating an induced null space, which intersects with the PLS model null space generated by the redundancy in the **Q** loadings due to the differences in the ranks of the **X** and **Y** matrices. The soft constraint on the Hotelling's T^2 proposed in eq 13 allows for the optimizer to move the solution along this null space toward the origin of the latent space so as to find a solution that satisfies the given constraints in the range of the historical data.

In Table 6 the solution obtained applying the procedure described for Method 1 (Section 3.2) by substituting the direct model inversion with the optimization framework in eq 13 is shown, when the equality constraint is set for ΔCompactibility ($\hat{\mathbf{y}}^{\mathrm{DES(4)}}$). Physical boundaries were specified for LOD and oversize percentage which were assigned to vary between 0 and 100, while $D90/D10$ and the growth ratio were set to be greater than 1. The solution is presented for the cases in which the soft constraint (SC) on the Hotelling's T^2 is considered ($g = 10^{-4}$) or not ($g = 0$) in the optimization formulation.

As can be seen, both of the obtained solutions satisfy the equality constraint on ΔCompactibility but are completely different from $\hat{\mathbf{y}}^{\mathrm{DES(4)}}$ in Table 4, in particular with respect to LOD and oversize percentage, which are found to be at their relevant boundaries. Note that the Hotelling's T^2 statistic for the solution obtained without considering the soft constraint on T^2 is above the 95% historical confidence limit represented in Figure 3. Including the soft constraint on the Hotelling's T^2 aids the optimizer to find a solution closer to the origin of the model space (and thus to the historical product profiles), and approximately lying on the model space ($\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}} \approx 10^{-6}$, second row of Table 6).

Two important considerations on the solution with the soft constraint on T^2 should be remarked.

First, it can be seen that the addition of the soft constraint into the objective function penalizes the minimization of $\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}}$, which is slightly different from zero. This is due to the fact that in this case the induced null space is actually a *pseudonull space*,[3] that is, moving the solution along it (while keeping fixed ΔCompactibility) does not ensure that a solution belonging to the model space is obtained, and slight deviations may therefore occur ($\mathrm{SPE}_{\mathbf{y}^{\mathrm{DES}}} \neq 0$). As can be seen, the decrease in T^2 due to the presence of the soft constraint is limited, as the boundaries specified for some of the variables do not allow moving the solution further toward the origin of the historical data model plane.

Second, note that in general the largest differences between the two solutions reported in Table 6 are mainly due to the growth ratio and to Δ*F*lodex. As can be seen from the first two plots of Figure 3, these variables scarcely affect the first two LVs of the model, which instead are those that better describe ΔCompactibility, while they are the most significant on the third LV.

Table 8. Method 2. Estimations of the Product Profiles $\mathbf{y}^{\mathrm{NEW}}$ Obtained at Each Iteration of the Procedure Described in section 3.3

	LOD (%)	oversize (%)	ΔFlodex (mm)	ΔCompactibility (KPa/MPa)	D[3,2] (μm)	$D$90/$D$10	growth ratio	T^2	SPE
$\mathbf{y}^{\mathrm{NEW}(1)}$	$3^{\#}$	$0^{\#}$	$20^{\#}$	$5^{\#}$	$400^{\#}$	$2.5^{\#}$	$8^{\#}$	8.8	7.3
$\mathbf{y}^{\mathrm{NEW}(2)}$	$3^{\#}$	$0^{\#}$	$20^{\#}$	$\underline{3}$	$400^{\#}$	$2.5^{\#}$	$8^{\#}$	3.0	4.4
$\mathbf{y}^{\mathrm{NEW}(3)}$	$3^{\#}$	$0^{\#}$	$20^{\#}$	$\underline{3}$	$\underline{29}$	$2.5^{\#}$	$8^{\#}$	6.7	2.9
$\mathbf{y}^{\mathrm{NEW}(4)}$	$3^{\#}$	$0^{\#}$	$20^{\#}$	$\underline{5.4}$	$\underline{3}$	$\underline{42.7}$	$8^{\#}$	49.9	0

This means that they are the most important variables on the induced null space, namely those which undergo the highest variations by moving the solution along it.

4.3. Method 2: Results.

In Table 7 the results obtained from the application of Method 2 (eq 13) are shown in terms of suggested new product quality sets $\mathbf{y}^{\mathrm{NEW}}$. The procedure has been applied specifying an additional constraint for ΔCompactibility, which was asked to be greater than 3 kPa/MPa (which represents a limit condition considering the available historical data set), while LOD and oversize percentage were assigned to vary between 0 and 100, and $D90/D10$ and the growth ratio were set to be greater than 1 (physical boundaries). Furthermore, with reference to section 3.3, ε was set to 10^{-6} in order to obtain a new product quality set $\mathbf{y}^{\mathrm{NEW}}$ that feasibly lies onto the model space. Results are reported for two cases: in the first case the soft constraint on the Hotelling's T^2 in eq 13 was not considered (i.e., $g = 0$), while in the second it was included ($g = 10^{-4}$). For both cases in Table 7, the values of $\mathbf{y}^{\mathrm{NEW}}$ (which according to the procedure can be kept equal to the ones specified in the original desired set $\mathbf{y}^{\mathrm{DES}}$) are indicated with the hashmark (#) superscript. In the last two columns, the values of the T^2 and SPE statistics for $\mathbf{y}^{\mathrm{NEW}}$ are reported, too.

As can be seen from the first row of Table 7 (no soft constraint on T^2), four of the seven product properties are maintained equal to the ones specified in $\mathbf{y}^{\mathrm{DES}}$, namely LOD, the percentage of oversize granules, ΔFlodex, and the growth ratio, while the solution $\mathbf{y}^{\mathrm{NEW}}$ can be considered lying onto the model space ($\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}}} \approx 0$). This means that the user can at most keep these values equal to the corresponding ones in $\mathbf{y}^{\mathrm{DES}}$ to obtain a set $\mathbf{y}^{\mathrm{NEW}}$ which belongs to the model space. In other words, Table 7 shows that to obtain a product that has the four indicated values assigned, the values of the other properties have to be those reported in the first row of Table 7 for the product to adhere to the historical product property covariance structure. Indeed $\mathbf{y}^{\mathrm{NEW}}$ represents the best trade-off between the original target quality profile $\mathbf{y}^{\mathrm{DES}}$ and the model requirements. Note that the T^2 value (49.9) indicates that an extrapolated solution above the T^2 confidence limits is eventually obtained (Figure 4). In general, this may not be a problem since, when applying LVRM inversion, this allows moving the solution along the null space in order to limit extrapolations in the input variable space, even if the desired product profile is out of the range of the historical products.[10]

The procedure described in Method 2 considering a soft constraint in T^2 was also applied. Note that, as discussed in section 3.1, an induced null space (depending on the subset of $\mathbf{y}^{\mathrm{DES}}$ being assigned) may be generated in this case, too. The second row of Table 7 shows the solution obtained considering the soft constraint on T^2 in eq 13. As can be seen, none of the variables keeps its value equal to those specified in $\mathbf{y}^{\mathrm{DES}}$. Moreover, note that the value of ΔCompactibility hits the specified inequality constraint (3 kPa/MPa), while the T^2 of the solution is significantly decreased, compared to the previous case (2.1 versus 49.9). In practice, in order to keep the solution onto the model hyperplane by limiting the corresponding Hotelling's T^2 and by satisfying the provided constraint, the procedure has to remove all the equality constraints on $\mathbf{y}^{\mathrm{DES}}$ by estimating a completely new product quality set $\mathbf{y}^{\mathrm{NEW}}$, in which only the inequality constraint on ΔCompactibility is satisfied.

To clarify the iterative procedure on which Method 2 is based, in Table 8 the estimations of the product profiles per iteration are reported together with the corresponding values of T^2 and SPE for the first case presented in Table 7, in which the soft constraint on T^2 is not considered. In Figure 5 the plots of the contributions of each variable to SPE ($\mathrm{SPE}_{\mathrm{CONT}}$) are shown for each iteration.

In Table 8, the variables for which the corresponding constraint is removed at each iteration are underlined, while those which the procedure keeps equal to the ones specified in the original desired set $\mathbf{y}^{\mathrm{DES}}$ are indicated with the # superscript. From the combined analysis of the plots of Figure 5 and the results in Table 8, a deeper insight on the algorithmic procedure can be obtained.

From the projection at the first iteration of $\mathbf{y}^{\mathrm{NEW}(1)} = \mathbf{y}^{\mathrm{DES}}$ onto the $\mathbf{Q}$ loadings of the PLS model, the contributions $\mathrm{SPE}_{\mathrm{CONT}}^{(1)}$ are calculated (Figure 5a). It can be observed that ΔCompactibility is the property most contributing to SPE. Therefore, the corresponding equality constraint is removed, and the optimization problem in eq 13 is solved to find the new set $\mathbf{y}^{\mathrm{NEW}}$ ($\mathbf{y}^{\mathrm{NEW}(2)}$ in Table 8), in which the value of ΔCompactibility satisfies the specified inequality constraint for it. It results that $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(2)}} = 4.4$, which is still above the threshold ε. In Figure 5b the contributions to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(2)}}$ are reported ($\mathrm{SPE}_{\mathrm{CONT}}^{(2)}$). It can be noted that the highest $\mathrm{SPE}_{\mathrm{CONT}}^{(2)}$ is still due to ΔCompactibility. However, the value of ΔCompactibility, whose equality constraint has already been removed, is kept fixed by the inequality constraint. Given that this inequality constraint cannot be relaxed due to the product requirements, in the next iteration, the optimization problem in eq 13 is solved removing the equality constraints on both ΔCompactibility and D[3,2], which is the variable with the second highest contribution to $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(2)}}$. The new set $\mathbf{y}^{\mathrm{NEW}(3)}$ (Table 8) presents $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(3)}} = 2.9$, still above ε; $\mathrm{SPE}_{\mathrm{CONT}}^{(3)}$ (Figure 5c) highlights that the model mismatch is mainly due to $D90/D10$. Accordingly, the optimization problem is solved again, removing the constraint on this variable, too. Finally, solution $\mathbf{y}^{\mathrm{NEW}(4)}$ exhibits $\mathrm{SPE}_{\mathbf{y}^{\mathrm{NEW}(4)}} \approx 0$ and thus it represents the optimal solution (Table 7; case without soft constraint on T^2).

By analyzing the iterative solution through Figure 5 and the values of SPE at each iteration, it can be noted that in this case a solution inside the 95% SPE confidence limit (Figure 4) is obtained by simply removing the equality constraint on ΔCompactibility ($\mathbf{y}^{\mathrm{NEW}(2)}$ in Table 8). The procedure could therefore have stopped at the first iteration. However, to allow for the inclusion of the hard constraints for $\hat{\mathbf{y}}^{\mathrm{DES}}$ in the LVRM inversion procedures (namely $\hat{\mathbf{y}}^{\mathrm{DES}} = \mathbf{Q}\boldsymbol{\tau}$), the methods were asked to find a solution very close to the model space ($\varepsilon \to 0$).

Note that in general Method 1 and Method 2 return different information. The first method provides a general perspective on how the assignment of a product specification affects the other variables according to the historical knowledge. The second method provides an optimal solution as a trade-off between the desired product quality variables and the need to fulfill the

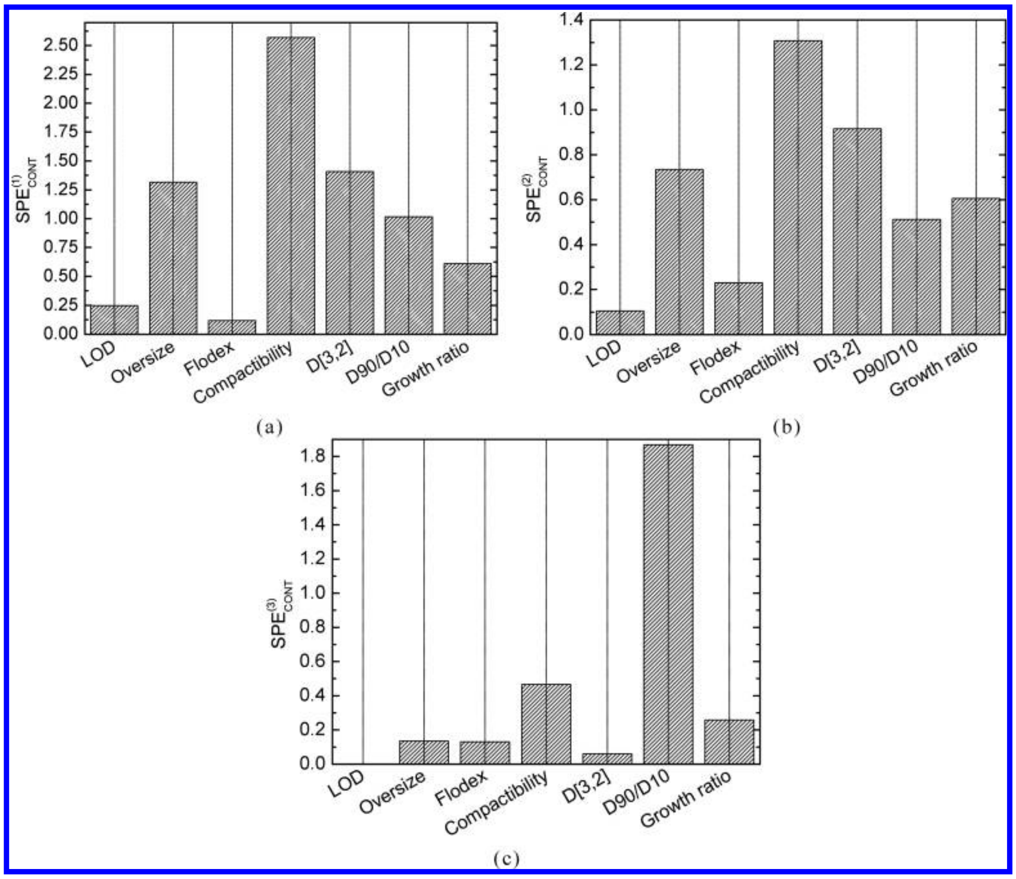

Figure 5. Method 2. Contribution plots obtained during the procedure iterations to calculate $\mathbf{y}^{\mathrm{NEW}}$: (a) 1st iteration; (b) 2nd iteration; (c) 3rd iteration.

relationships between product variables obtained from the historical data and represented by the **Q** loadings of the PLS model. However, it must be noted that Method 1 is more susceptible to uncertainty due to the calculation of the inverse of the $\mathbf{Q}^{(i)^{\mathrm{T}}}\mathbf{Q}^{(i)}$ matrix. Depending on the variable which is assigned for $\hat{\mathbf{y}}^{\mathrm{DES}}$ or on the number of LVs selected to build the PLS model, matrix $\mathbf{Q}^{(i)^{\mathrm{T}}}\mathbf{Q}^{(i)}$ may be ill-conditioned due to variable correlation. The ill-conditioning and the presence of noise in the measurements, which masks the effective rank of the **Y** space, could result in poor estimations of the solution scores using eq 10. For this reason, as shown in section 4.2, Method 1 can be applied by substituting the direct inversion of the model with an optimization framework, which allows moving the solution along the (induced) null space to find the minimum Mahalanobis distance (i.e., with minimum Hotelling's T^2) solution.[3,4]

Finally consider that it may occur that none of the Methods returns a solution if the design of a product with very different properties from the ones in historical data set is required.

5. CONCLUSIONS AND FUTURE WORK

Latent variable regression model (LVRM) inversion has long been demonstrated to be an effective tool to support product and process design. In this paper we proposed a methodology to exploit the covariance structure of the historical data used to build the LVRM so as to guide the selection of a target attribute profile ($\mathbf{y}^{\mathrm{DES}}$) with the same covariance structure as the matrix **Y** of the responses. The method exploits the **Q** loadings of the PLS model to suggest possible sets of target profiles as trade-offs between the model requirements and the desired product characteristics. Depending on the structure of the latent space of **Y**, the loadings from a PCA model on **Y** can be used as an alternative.

Two different procedures have been presented. In the first one, product quality features are assigned one at a time, and an approach based on model inversion of the model is used to estimate the other variables. In the second procedure, an algorithm is used to iteratively select and remove the constraints of the variables that are found to be most responsible for the model mismatch. An optimization problem is solved to calculate their values according to the loadings of the **Y** space. We demonstrated how in both these methods an induced null space may be generated depending on the assigned variables, in which different solutions, all satisfying the given constraints can be found.

The proposed approaches have been applied to a real case-study concerning a high-shear wet granulation process. The methods have shown their effectiveness in assessing the feasibility of a new product, and in suggesting a reliable and physically sound product quality profile.

Note that the methodology relies on empirical data only and, accordingly, the feasibility of a product design is exclusively defined on the basis of the previous knowledge. However, we believe that the methodology may prove to be a useful tool not only to understand if an empirical model is appropriate for inversion, but also to assess the potential quality profiles of completely new products since the very first developmental stages.

Future work may involve the investigation of using a "slack variable" approach instead of a weighted term in the objective function to address the stiffness induced by adding hard equality constraints to the problem; such an approach would need to investigate the theoretical considerations of having a relaxed constraint with a threshold and any potential null-spaces acting together. Additional avenues for future work are to incorporate the uncertainty in the model parameters into the objective function (or the constraints).

APPENDIX A. ON THE RECONSTRUCTION OF Y^{DES}

As seen above, to reconstruct $\hat{\mathbf{y}}^{DES}$ a model describing the covariance structure of the historical data in **Y** is needed. In general, the covariance structure of the historical data can be optimally described by a principal component analysis (PCA[18]) on the historical product data set **Y**, apart from the **Q** loadings of the PLS model between **X** and **Y** (eq 2). Thus $\hat{\mathbf{y}}^{DES}$ could feasibly be reconstructed either exploiting the PCA or the PLS model loadings.

However, the covariance structure described by the PCA model on **Y** could potentially not be as the one described by the **Q** loadings of the PLS model. Considering the different objectives of the two techniques, reconstructing $\hat{\mathbf{y}}^{DES}$ through the PCA model forces the target product quality $\mathbf{y}^{DES}$ to respect the correlation between the quality variables of the historical products in **Y**. Differently, the **Q** loadings of the PLS model are calculated so that the LVs of **Y** are rotated to maximize their covariance with the LVs of **X**. Thus the systematic variability of **Y** which is related to that in **X** is considered for the reconstruction of $\hat{\mathbf{y}}^{DES}$.

Given that the objective of this paper is that of proposing a procedure which allows an imposition of the hard constraint for $\mathbf{y}^{DES}$ in the optimization formulation of the model inversion problem, the reconstruction of $\hat{\mathbf{y}}^{DES}$ has been based on the **Q** loadings of the PLS model. The proposed methodologies can however be applied with no modifications to the case in which the PCA loadings are used to reconstruct $\hat{\mathbf{y}}^{DES}$.

APPENDIX B. INTERPRETATION OF THE LOADING PLOTS OF THE WET GRANULATION CASE STUDY

The 1st LV, which accounts for ~32% of the total variance in **Y** (Table 3), is mainly driven by the particle size distribution (PSD) variables, namely $D[3,2]$ and $D90/D10$ (top plot of Figure 3). In particular, $D[3,2]$ and $D90/D10$ are opposite, meaning that granulated materials in the database with high PSD (high $D[3,2]$) have usually narrower PSD (low $D90/D10$), compared to the mean of the data. This seems to affect the percentage of oversize granules, which is directly correlated with $D[3,2]$ and the difference in compactibility compared to the raw material (ΔCompactibility). Namely, it is expected that products with high PSD mean, have larger percentages of oversize granules compared to materials with lower PSD means, and this seems to slightly affect the compactibility difference of the granulated product with the raw material, which is expected to be lower.

The 2nd LV, which explains ~23% of the total variance of the data, is affected mainly by the moisture content loss upon drying (LOD) of the materials after the granulation (second plot of Figure 3). From the analysis of the variable loadings it can be argued that materials which have a higher loss of water upon drying are in general those which are less grown and result in products with lower compactibility properties compared to the raw materials (ΔCompactibility).

The 3rd LV (third plot of Figure 3) explains ~8% of the total variance of the data and is mainly driven by the flow properties of the products (ΔFlodex). In particular it seems that products which are more flowable than the corresponding raw materials are also those which have had higher growth ratio and LOD. This is somehow expected since one of the objectives of granulation is that of increasing the flow properties of the processed materials, by enlarging their size. Moreover, moisture (inversely related to LOD) can act as a binding, making the granules more cohesive and thus less flowable.[19] The 4th LV seems to be less significant than the other ones in explaining the systematic variability of **Y**, as reported in Table 3 (1.58%). This can be also noticed from the low value of the loadings in the last plot of Figure 3.

AUTHOR INFORMATION

Corresponding Author

*E-mail: sal.garcia@pfizer.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

E.T. gratefully acknowledges "Fondazione Ing. Aldo Gini" for the financial support. The authors want also to thank Prof. Massimiliano Barolo for his help and precious discussions to improve the value of the work and the quality of the manuscript.

REFERENCES

(1) Jaeckle, C. M.; MacGregor, J. F. Product design through multivariate statistical analysis of process data. *AIChE J.* **1998**, *44*, 1105.

(2) Jaeckle, C. M.; MacGregor, J. F. Industrial applications of product design through the inversion of latent variable models. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 199.

(3) García-Muñoz, S.; Kourti, T.; MacGregor, J. F.; Apruzzese, F.; Champagne, M. Optimization of batch operating policies. Part I. Handling multiple solutions. *Ind. Eng. Chem. Res.* **2006**, *45*, 7856.

(4) García-Muñoz, S.; MacGregor, J. F.; Neogi, D.; Letshaw, B. E.; Mehta, S. Optimization of batch operating policies. Part II. Incorporating process constraints and industrial applications. *Ind. Eng. Chem. Res.* **2008**, *47*, 4202.

(5) Flores-Cerrillo, J.; MacGregor, J. F. Control of batch product quality by trajectory manipulation using latent variable models. *J. Process. Control* **2004**, *14*, 539.

(6) Flores-Cerrillo, J.; MacGregor, J. F. Latent variable MPC for trajectory tracking in batch processes. *J. Process. Control* **2005**, *15*, 651.

(7) García-Muñoz, S.; Dolph, S.; Ward, H. W., III Handling uncertainty in the establishment of a design space for the manufacture of a pharmaceutical product. *Comput. Chem. Eng.* **2010**, *34*, 1098.

(8) Yacoub, F.; MacGregor, J. F. Product optimization and control in the latent variable space of nonlinear PLS models. *Chemom. Intel. Lab. Sys.* **2004**, *70*, 63.

(9) Muteki, K.; MacGregor, J. F.; Ueda, T. Rapid development of new polymer blends: The optimal selection of materials and blend ratios. *Ind. Eng. Chem. Res.* **2006**, *45*, 4653.

(10) Tomba, E.; Barolo, M.; García-Muñoz, S. General framework for latent variable model inversion for the design and manufacturing of new products. *Ind. Eng. Chem. Res.* **2012**, *51*, 12886.

(11) Höskuldsson, A. PLS regression methods. *J. Chemom.* **1988**, *2*, 211.

(12) Burnham, A. J.; Viveros, R.; MacGregor, J. F. Frameworks for latent variable multivariate regression. *J. Chemom.* **1996**, *10*, 31.

(13) Macgregor, J. F.; Kourti, T. Statistical process control of multivariable processes. *Control Eng. Practice.* **1995**, 3, 3.

(14) Arteaga, F.; Ferrer, A. Dealing with missing data in MSPC: Several methods, different interpretations, some examples. *J. Chemom.* **2002**, *16*, 408.

(15) Nelson, P. R. C.; Taylor, P. A.; MacGregor, J. F. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemom. Intel. Lab. Sys.* **1996**, *35*, 45.

(16) Kerrigan, E. C.; Maciejowski, J. M. Soft constraints and exact penalty functions in model predictive control. In: Proc. UKACC

International Conference (Control 2000): Cambridge (U.K.), 4–7 September 2000.

(17) Vemavarapu, C.; Surapanemi, M.; Hussain, M.; Badawy, S. Role of drug substance material properties in the processability and performance of a wet granulated product. *Int. J. Pharm.* **2009**, *374*, 96.

(18) Jackson, J. E. *A User's Guide to Principal Components*; John Wiley & Sons, Inc.: New York, 1991.

(19) Emery, E.; Oliver, J.; Pugsley, T.; Sharma, J.; Zhou, J. Flowability of moist pharmaceutical powders. *Powder Technol.* **2009**, *189*, 409.