Computers and Chemical Engineering 115 (2018) 309–322

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

A novel and systematic approach to identify the design space of pharmaceutical processes

Gabriele Bano[a], Zilong Wang[b], Pierantonio Facco[a], Fabrizio Bezzo[a], Massimiliano Barolo[a], Marianthi Ierapetritou[b,*]

[a] CAPE-Lab – Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, via Marzolo 9, 35131 Padova, PD, Italy

[b] Department of Chemical and Biochemical Engineering, Rutgers University, 98 Brett Road, Piscataway, NJ 08854, USA

ARTICLE INFO

Article history:
Received 21 November 2017
Revised 13 April 2018
Accepted 18 April 2018
Available online 22 April 2018

Keywords:
Feasibility analysis
Partial least-squares regression
Design space
Continuous pharmaceutical manufacturing

ABSTRACT

Feasibility analysis is a mathematical technique that can be used to assist the identification of the design space (DS) of a pharmaceutical process, given the availability of a process model. One of its main drawbacks is that it suffers from the curse of dimensionality, i.e. simulations can potentially become computationally extremely expensive and very cumbersome when the number of input factors is large. Additionally, giving a graphical and compact representation of the high-dimensional design space is difficult. In this study, we propose a novel and systematic methodology to exploit partial least-squares (PLS) regression modelling to reduce the dimensionality of a feasibility problem. We use PLS to obtain a linear transformation between the original multidimensional input space and a lower dimensional latent space. We then apply a Radial Basis Function (RBF) adaptive sampling feasibility analysis on this lower dimensional space to identify the feasible region of the process. We assess the accuracy and robustness of the results with three metrics, and we critically discuss the criteria that should be adopted for the choice of the number of latent variables. The performance of the methodology is tested on three simulated case studies, one of which involving the continuous direct compaction of a pharmaceutical powder. In all case studies, the methodology shows to be effective in reducing the computational burden while maintaining an accurate and robust identification of the design space.

1. Introduction

The design space (DS) of a pharmaceutical process is defined as the multidimensional combination and interaction of input variables (e.g. material attributes) and process parameters that have been demonstrated to provide assurance of quality (ICH, 2009). A modeling technique that can be used to assist a DS identification exercise is feasibility analysis (Boukouvala et al., 2010; Wang and Ierapetritou, 2017).

The aim of feasibility analysis (Halemane and Grossman, 1983) is to quantify the capability of a process design to be feasibly operated over the whole domain of input factors, thus including raw material properties and process parameters. The final objective is to determine the multivariate region of the input domain within which the process is considered to be feasible. Hence, the concept of feasible region is strictly related to the one of DS for pharmaceutical processes.

Different mathematical approaches can be used to identify the feasible region. When the model of the process is computationally inexpensive, the feasible region can be directly determined from the model itself (Prpich et al., 2010; Close et al., 2014). When the original model is computationally expensive or the computation of its derivatives is cumbersome, surrogate-based approaches can be used (Boukouvala et al., 2010; Grossmann et al., al.,2014; Wang and Ierapetritou, 2017; Wang et al., 2017). The underlying idea behind these methods is to build a surrogate as a computationally cheap and reliable approximation of the original model, and use this surrogate to identify the feasible region. In this regard, different types of surrogate models have been used (Goyal and Ierapetritou, 2002; Banerjee et al., 2010; Boukouvala et al., 2010; Rogers and Ierapetritou, 2015). Recently, Wang and Ierapetritou (2017) proposed a radial basis function (RBF) adaptive sampling approach that outperforms all the other surrogate-based approaches for low dimensional test problems. The adaptive sampling is a technique that was first proposed in the optimization literature (Jones et al., 1998). In surrogate-based feasibility analysis, it is used to maximize the potential of the sampling budget and to simultaneously explore re-

* Corresponding author.
E-mail address: marianth@soe.rutgers.edu (M. Ierapetritou).

https://doi.org/10.1016/j.compchemeng.2018.04.021
0098-1354/

List of symbols

A	number of latent variables
$\mathbf{d}$	vector collecting all the design variables
$\mathbf{E_X}$	**X-** residuals
$\mathbf{E_Y}$	**Y-** residuals
$EI_{feas}(\mathbf{x})$	modified expected improvement function
$f_j(\mathbf{d}, \mathbf{x})$	j-th quality constraint
I	number of historical samples
J	number of quality constraints
L_{PLS}	linear tranformation of variables through the PLS model
M	number of responses (outputs)
$\mathbf{P}$	**X-** loadings matrix
Q	number of input factors
$\mathbf{Q}$	**Y-** loadings matrix
R^2_X	R^2 statistic for the **X**-variability
$R^2_{X,cum}$	cumulative R^2 statistic for the **X**-variability
R^2_y	R^2 statistic for the **y**-variability
$R^2_{y,cum}$	cumulative R^2 statistic for the **y**-variability
$\mathbf{t}^f$	feasible point on the latent space
$\mathbf{T}$	**X-** scores matrix
$\mathbf{x}$	vector that collects all the input factors
$\mathbf{x}^f$	feasible point on the original input space
$\mathbf{x}^*$	unsampled point
$\mathbf{X}$	historical input matrix
$\mathbf{Y}$	historical output matrix
$\hat{y}$	surrogate prediction
$y_n(\mathbf{x})$	value of the original function at an unsampled point
$\mathbf{W}$	weight matrix

Greek letters

ξ	basis function
$1/\mu_n(\mathbf{x}^*)$	error indicator
$\psi(\mathbf{x})$	feasibility function

Acronyms

DoE	design of experiments
DS	design space
EI	expected improvement
KSA	Kennard-Stone's algorithm
LV	latent variable
LVM	latent variable model
PCA	principal component analysis
PLS	partial least-squares
RBF	radial-basis function

gions of the original input domain that are close to the boundary of the feasible region and less explored regions.

Although feasibility analysis can be a valuable tool to identify the feasible region of a manufacturing process, it suffers from one main limitation, namely the curse of dimensionality (Shan and Wang, 2010). When a large number of input factors is involved, as it is often the case with large and complex integrated flowsheet models, the computational cost of feasibility analysis has a potential to increase significantly. The solution of a feasibility analysis problem in this scenario can be extremely complicated if not impossible to solve. When the number of input factors is large, the computational burden could be so high or the results so inaccurate as to effectively preclude the application of this methodology. Moreover, a visualization problem of the feasible region in high dimensions arises, given the impossibility of graphically representing a N- dimensional space ($N > 3$).

A recent work of Wang et al. (2017) tried to solve this problem by transforming a multidimensional feasibility analysis problem into a series of disjoint 2-dimensional problems, and presenting the results as a matrix of 2-dimensional feasibility contour plots. However, this approach does not account for the multivariate correlation between the original input factors and is thus incomplete.

The input factors of a pharmaceutical process are often correlated to each other (Tomba et al., 2013) and have a different impact on the process output (e.g. product quality; Saltelli et al., 2010). In most common situations, not all the combinations of the original input factors have a strong effect on the output, i.e. some "driving forces" can be identified that predominantly affect the responses (Jaeckle and McGregor, 2000). A class of statistical models that can identify these underlying driving forces are latent variable models (LVMs). In particular, partial least-squares (PLS) regression (Geladi and Kowalski, 1986; Wold et al., 1983) is a multivariate latent variable technique that can be used to capture the variability of the input and output spaces by means of few meaningful variables, called latent variables, thus reducing the problem dimensionality. The latent variables are chosen by performing a simultaneous decomposition of the input and output space, such that these variables explain as much as possible of the covariance between the two spaces. Stated differently, PLS identifies linear combinations of the original input factors that best describe the correlation between the input factors *and* their effect on the model responses. PLS can be used by itself as a black-box data-driven modelling technique to assist a DS identification exercise when a historical dataset of the process under investigation is available (Facco et al., 2015; Bano et al., 2017). However, when a historical dataset is not available or the process behavior cannot be captured by a linear regression model, the ability of PLS to "compress" the original multidimensional input and output spaces (Mevik et al., 2004) can be coupled with more sophisticated modelling techniques to assist a DS identification exercise.

The aim of this work is to overcome the curse of dimensionality problem of feasibility analysis in DS identification by exploiting the ability of PLS to reduce the input space dimensionality. We use a PLS model to perform a linear transformation from the original multidimensional input space to a lower dimensional latent space. Based on the PLS diagnostics, we identify and critically discuss different scenarios in which the proposed methodology is deemed to be profitable.

We perform the RBF-based adaptive sampling feasibility analysis on the latent space and we assess the accuracy and robustness of the results with three appropriate metrics. We test the performance of the methodology on three simulated case studies, two of them involving two-and three unit multidimensional test problems, and one involving a continuous direct compaction of a pharmaceutical powder. In all case studies, we prove the ability of the methodology to give accurate and robust results by simultaneously reducing the computational burden for the feasibility analysis problem.

The rest of this article is organized as follows. We first introduce PLS and RBF-based adaptive sampling feasibility analysis in Section 2. We then introduce the proposed methodology to couple PLS and feasibility analysis for the identification of the feasible region of a manufacturing process in Section 3. Section 4 collects the case studies considered in this work and Section 5 shows the results obtained with the proposed methodology.