VARIABLE REDUCTION FOR SURROGATE MODELLING

J. Straus[1] and S. Skogestad*[1]

[1]Department of Chemical Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Abstract

In this paper, we present a three-step procedure for the reduction of independent variables u for surrogate modelling. First, the linear material balances are introduced to reduce the number of surrogate models which need to be fit. Second, partial least square (PLS) regression of a sampled space is performed to obtain new variables (components) and third, the new components are used as input variables for the fitting of a nonlinear surrogate model. The application of PLS reduces the number of independent variables through the introduction of linear combinations of the original independent variables u. The proposed procedure is applied to two examples, the first describes a simple pipe model in which the minimum number of new independent variables u' is known and which hence serves as a proof of concept. The second examples considers the reaction section of the ammonia synthesis gas loop for integrated submodels. In both examples, it is possible to reduce the number of independent variables by at least a factor of 2 while maintaining accuracy.

Keywords

Surrogate modelling, PLS regression, Independent variable reduction.

Introduction

Surrogate models are frequently used to incorporate complicated models into numerical demanding simulations and are defined as auxiliary models fitted to generated data. Their applications range from the production optimization of oil gas fields (Grimstad et al., 2016; Foss et al., 2015) and optimization of CFD simulations (Badhurshah and Samad, 2015) to modelling of chemical process (Cozad et al., 2014, 2015; Caballero and Grossmann, 2008). The structure of surrogate models can range from the simple table look up method, in which generated data is stored in arrays and extracted when needed, to splines, Kriging models, artificial neural networks, and combinations of several different basis functions.

The optimization of integrated chemical plants using commercial steady-state flowsheet simulators, like Aspen Plus®, Aspen Hysys®, SimSci PRO/II, or UniSim Design Suite is in general difficult directly due to the sequential-modular approach to solve the flowsheet, in which each unit operation is solved sequentially (Biegler et al., 1997). In the situation of several (nested) recycle streams, convergence issues arise leading to cases, in which it is even not possible to solve the flowsheet. Furthermore, recycle loops can introduce numerical inaccuracy. Therefore, the application of surrogate models for subsystems and hence splitting the main recycle streams seems advantageous.

Previously, we have proposed (Straus and Skogestad, 2016) a methodology for the optimization of integrated process which involves splitting the complete flowsheets into several submodels, fit surrogate models to these submodels and subsequently combining the surrogate models into a system of non-linear equations which can be optimized. However, due to the connection variables, the number of independent variables (n_u) is generally quite high. This can lead to problems caused by the dimensional "curse" of surrogate models if regular grids are used, the exponential dependency of the surrogate model fitting to n_u as given in Eq. (1).

$$n_{RG} = 2^{n_u} \tag{1}$$

*To whom all correspondence should be addressed *skoge@ntnu.no*

Hence, it is important to keep n_u small, preferably less than four (Grimstad et al., 2016).

The reduction of n_u can be based on process knowledge, *e.g.* neglecting independent variables that are known to have a negligible influence on the output variables. Another possibility is to introduce new independent variables u', which can be, among others, derived *via* partial least square (PLS) regression. PLS regression is a linear regression tool in which the predicted and observable variables are projected into a new space through the introduction of components. It was developed by Wold et al. (1983) to solve the multivariate calibration problem in the case of chemometrics. In this problem, the number of sampling points is less than the number of independent variables, *i.e.* the number of varied concentrations is smaller than the number of measured frequencies and an optimal combination of measurements for concentration regression has to be found. Similarly, it was applied in the analysis of genomic data (Boulesteix and Strimmer, 2007). Based on the mentioned previous applications of PLS regression, it seems to be a reasonable tool for the reduction of the number of independent variables. It has to be noted, that the procedure itself is not limited to partial least square regression, but can also utilize for example dimensionless numbers as well.

Procedure for Dimension Reduction

The overall procedure to reduce n_u consists of in total three steps; introduction of the linear material balance relationships, independent variable dimension reduction, and fitting of the surrogate model to the new independent variables u'. It will be explained in the following subsections. In addition, this procedure allows the use of a new model structure which is visualized below in figure 1. This methodology requires the initial sampling of a certain number of points n_p to perform the PLS regression which will be in the following denoted as $U \epsilon \mathbb{R}^{n_p \times n_u}$.

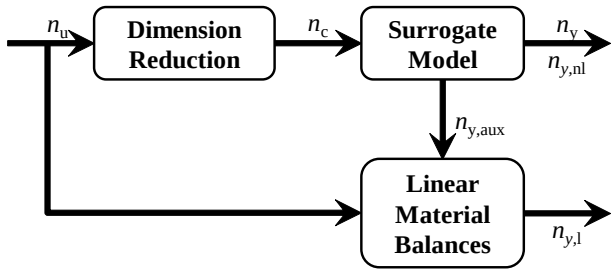

Figure 1. Structure of the proposed new model structure.

We propose to incorporate the corner points of a regular grid of the sampling space in order to not extrapolate data within the investigated sampling space and add additional points through Latin hypercube sampling or orthogonal sampling to guarantee a proper distribution of the points.

Definition of Linear Relationships

Linear input-output relationships can be always defined for mass balances and in certain cases for the energy and force balances. This can be reasoned by the knowledge of the flowsheet topology in the case of production optimization. However, the application of linear relationships may require the introduction of auxiliary variables y_{aux}. In the case of a reaction within the submodel, the extent of reaction ξ in combination with the stoichometric factor ν_i allows the reduction of the number of surrogate models to be fitted. If, on the other hand, a separation takes place in the submodel or a split is present, the mass balances can be introduced *via* a separation coefficient p. The introduction of linear relationships hence reduces the number of surrogate models which have to be fitted.

In addition, the introduction of linear mass balances results in mass consistency. If this step would not be conducted, the combination of surrogate models could lead to creation or removal of mass due to model inaccuracy rendering their application doubtful.

Dimension Reduction

As mentioned, the application of PLS regression yields as a result linear combinations of the initial independent variables, which represent the nonlinear output variables y_{nl} and/or y_{aux} for the given sampled data best. It is important to mention, that a PLS regression should be performed for each of the output variables y_{nl} and y_{aux} as otherwise components are chosen with a trade-off for fitting the output variables to the independent variables.

An additional advantage of the application of PLS regression is that it gives an overview about the influence of the independent variables u on the derived nonlinear output values y_{nl} and y_{aux}. This can be utilized for the addition of points to the sampling grid in the relevant direction, but will not be elaborated further. The linear combinations of the components defined are hereby independent of the total number of components. This means, that the linear combination of the first compo-

nent will be the same if $n_c = 1$ or $n_c = n_u$. Therefore, it is useful to perform the PLS regression directly for $n_c = n_u$ components and only use the first k components for the definition of the surrogate model in the subsequent fitting. Before applying PLS regression, it is additional advantages to perform variable transformations for the independent variables. If it is for example known, that the partial pressure of components or the total flow play a crucial role, it is useful to redefine the matrix for the sampled space U in terms of total flow Q and mole fractions x_i or partial pressures p_i.

The SIMPLS algorithm used by MATLAB for PLS regression is strongly depending on the scaling of the variables. Hence, it is crucial to scale the sampled space appropriately. If the scaling is not performed properly, the first component will point towards the space instead of capturing the true component. In the following, the standard score will be applied for scaling the sample space U which is defined as

$$U_{scaled} = (U - \mu_U) \circ / \sigma_U \tag{2}$$

where μ_U is the mean value and σ_U the standard deviation in the matrix U with respect to each of the independent variables u. Using the standard score, we scale the input matrix U in way that we assume the variance of each independent variable is equal. However, in cases where we would like to preserve the changes in the independent variables, the scaled matrix U_{scaled} can be further adjusted using a scaling matrix S_U, for example, corresponding to the percentage change in the sampling space.

Surrogate Model Fitting

The surrogate model is then fitted to the new independent variables $c = u'$ defined as linear combinations of the original independent variables u. The fitting of the surrogate model is an iterative procedure in which the number of components, n_c is increased until a fitting criteria is fulfilled. Alternatively, the explained variance per component in the response (y_{nl} and/or y_{aux}) can be utilized as a starting point. The type of surrogate model is not important for this procedure. For example, artificial neural networks, splines, Kriging models, or polynomials can be applied. However, due to the introduction of new independent variables, it is necessary that the surrogate model basis functions do not require a regular grid as a regular grid will not exist after variable transformation through PLS.

Algorithmic Approach

The above procedure can also be written as a pseudocode for the calculation of the surrogate models as shown below.

Algorithm 1 Procedure for independent variable reduction.

1: Define sampling grid A of the problem.
2: Sample training and validation space.
3: Define linear relationships if possible.
4: **for** $k = 1$ to $n_{y,nl} + n_{y,aux}$ **do**
5: Perform PLS regression with $n_{c,j} = n_u$.
6: **while** $\epsilon_j > threshold$ **do**
7: Fit surrogate model g' to $n_{c,j} = k$.
8: $x_{sm,j} = g'(n_{c,j})$
9: $\epsilon_j = \frac{|x_{val,j} - x_{sm,j}|}{x_{val,j}}$.
10: $k = k + 1$

Example 1: Pipe Model

The pipe model is used as a proof of concept model. The model gives the pressure drop over a pipe as a function of the independent variables inlet pressure p_{in}, temperature T_{in}, and component molar flows $\dot{N}_{i,in}$. The total number of independent variables n_u is hence given by $n_u = 2 + n_{chem}$ in which n_{chem} is the number of chemicals in the gas stream.

Model

The model itself consists of an isothermal pressure drop given in Eq. (3)

$$p_{in}^2 - p_{out}^2 = 4f\frac{L}{D}\frac{RT_{in}\bar{M}}{A^2}\dot{N}^2 \tag{3}$$

Based on step 1 in the procedure, we can introduce as linear balances the constant temperature assumption and the mass balances

$$T_{in} = T_{out} \tag{4}$$

$$\dot{N}_{i,in} = \dot{N}_{i,out} \quad \text{for } i = 1...n_{Chem} \tag{5}$$

This leaves as a nonlinear relationship the calculation of the outlet pressure. Hence, one surrogate model has to be defined. Simulations with 3, 5, and 8 chemicals are performed to demonstrate the procedure. The sampled space is given by a 2-point regular grid with an additional 100 (1000 and 5000 respectively for 5 and 8 chemicals) points defined as a Latin hypercube. This corresponds in each case to about 2.5 points in a regular grid. After performing the PLS regression, a 1-layer

cascade-forward neural network with 5 hidden neurons was fitted using the new independent variables defined *via* PLS regression and the performance of the surrogate model was evaluated with 10^4 points sampled as a Latin hypercube with the same bounds.

Results of the Reduction in Independent Variables

From Eq. (3), we can directly see that four independent variables, p_{in}, T_{in}, $\bar{M}$, and $\dot{N}$, are sufficient for the full characterization of the system and it is not necessary to know the exact composition of our gas stream as long as we now the average molar mass $\bar{M}$. As the PLS components are always taking into account the previous, unchanged linear combinations, it has to be noted, that a similar performance cannot be expected.

A PLS regression with 2, 3, and 4 components gives the results in figure 2. It can be seen that the number of variable reduction through PLS allows as little as 3 independent components. Increasing the number of components to 4 only marginally improves the performance of the surrogate model fitting. This is confirmed by the explained variance through PLS regression for the response variable p_{out}; from 2 to 3 components, it is increased from 77.71% to 99.56% whereas the increase to 4 components only has an influence on the explained variance in the predictor variable matrix U_{scaled}. Analogous results can be found in the case of 5 and 8 chemicals. The increased accuracy for 5 and 8 chemicals is given by the increased number of points the surrogate model is fitted to, as the regular grid for the initial independent variables u is exponentially increasing with the number of independent variables as shown in Eq. (1). Increasing the sampling space in the case of 3 chemicals to the same number as points as in the case of 8 chemicals results in similar relative errors, confirming this reasoning.

Example 2 - Reaction Section of the Ammonia Synthesis Loop

The reaction section of the ammonia synthesis gas loop is an example of an integrated process. The model consists of 2 reactor beds and is illustrated in figure 3. To exploit the produced heat and improve reactor utilization through shifting of the thermodynamic equilibrium, several heat exchangers are introduced preheating the feed to the first bed by cooling the effluent of both beds.

In this model, we have two nested recycle loops (M-R1-HEx2-M and M-R1-HEx2-R2-HEx3-M) and a third recycle loop in contact with the nested (HEx1-S-HEx3-HEx4-HEx1). Incorporating this model into a big flowsheet may lead to time-consuming flowsheet evaluations which makes it not useful for optimization. The number of independent variables $n_u = 10$ is given by the variables of the feed stream (7 variables: p_{in}, T_{in}, and $\dot{N}_{i,in}$) plus the two split ratios through the valve and heat exchanger 2 as well as the outlet temperature ($T_{Ref,4}$) of heat exchanger 4. The split ratio through heat exchanger 3 is defined *via* the aforementioned split ratios to maintain no mass accumulation in the split. $n_u = 10$ is generally considered much too high for surrogate modelling as it would for example in the case of B-splines only allow 3 points for the surrogate model design (Grimstad et al., 2016) corresponding to 59,049 flowsheet evaluations.

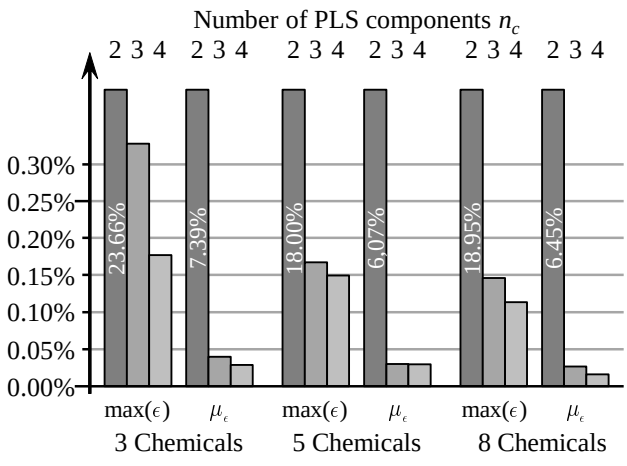

Figure 2. Relative error achieved after surrogate model fitting for the pipe model.

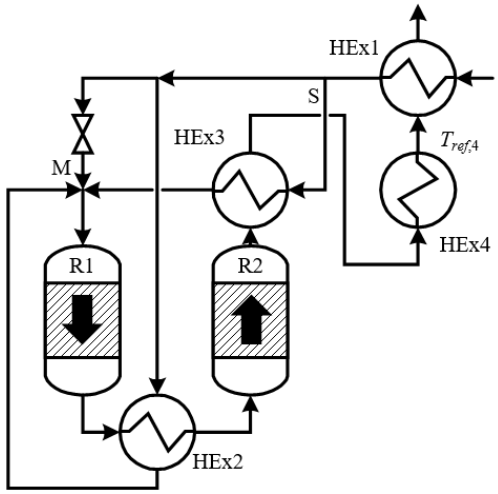

Figure 3. Flowsheet of the reaction section of the ammonia synthesis loop.

Hence, independent variable reduction is necessary and the outlined procedure will be applied.

Model

The flowsheet was modelled in MATLAB and compromises a non-linear system of equations with 282 states. The reactor beds are modelled as CSTR-cascades and the heat exchangers using the Number of Transfer Units Method. In step 1 of the proposed procedure, linear relationships for the mass balances are introduces using the extent of reaction ξ as

$$\dot{N}_{i,out} = \dot{N}_{i,in} + \nu_i \xi \tag{6}$$

This leaves nonlinear relationships for p_{out}, T_{out}, and ξ. Hence, 3 surrogate models have to be fitted in total. The sampled grid is given by a 2 point regular grid and 5000 additional points defined as a Latin hypercube. The fitted surrogate models are 3-layer cascade-forward neural networks with 2, 5, and 5 hidden neurons in the layers respectively. The resulting model is then validated with 10^5 points sampled as a Latin hypercube. It has to be mentioned, that the neural network structure was not optimized with respect to the different output variables y. In addition, the sampling space was chosen too small for the fitting of a non-linear model to a regular grid as it corresponds to 2.39 points for each independent variable.

Results of the Reduction in Independent Variables

Compared to the pipe model, it is this time not possible to define the minimum number of components ($n_{c,min}$) necessary to fit a surrogate model to accurately predict the outlet pressure p_{out}, the outlet temperature T_{out} and the extent of reaction ξ. In this situation, it is useful to start at a minimum value for the number of components of $n_c = 5$ and continue in a positive or negative reaction, depending on the fit of the surrogate model. From experience it is expected, that it can be beneficial to describe the problem in terms of a total flow Q_{in} and mole fractions $x_{i,in}$ for PLS regression instead of using the mole flows $\dot{N}_{i,in}$. In order to fulfill that the numbers of independent variables remain the same, one mole fraction has to be left out, in this case the mole fraction of one of the inerts methane or argon, as they are the least interesting.

The results for the outlet pressure p_{out} can be found in figure 4. From this figure, we see that the pressure drop over the system can be accurately describe by

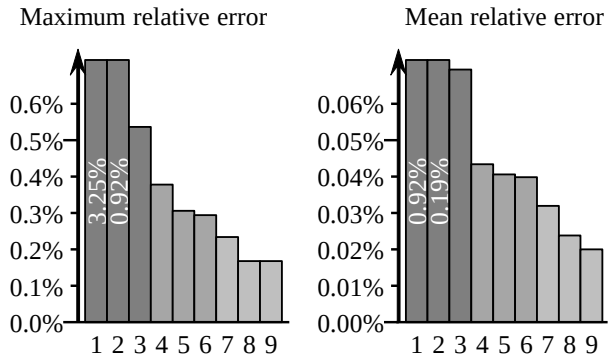

Figure 4. Relative error of the outlet pressure as a function of the number of PLS components n_c.

four or more components obtained *via* PLS regression. In absolute values, the maximum and mean error for four components are given by 0.2 bar and 0.02 bar respectively. Here, it is interesting to note that the explained variance in the response p_{out} is increasing from one to four components from 96.9% to 99.94%, which corresponds to the improved fit of the surrogate model shown in figure 4.

Similar to the outlet pressure p_{out}, the outlet temperature T_{out} can be adequately described with four or more PLS component as shown in figure 5. In general, the maximum and mean relative error is higher than in the case of the outlet pressure. However, the maximum and mean error is only 0.20 °C and 0.02 °C respectively. Analogous to p_{out}, a drastic improvement can be found by increasing the number of PLS components from 1 to 4. The improvement in the explained variance in the response T_{out} is increasing in these steps as well from 99.83% to 99.99% showing that the explained variance can be used for analyzing results, but not for prediction of the accuracy of the model fit. Otherwise, one would conclude that one component would be sufficient.

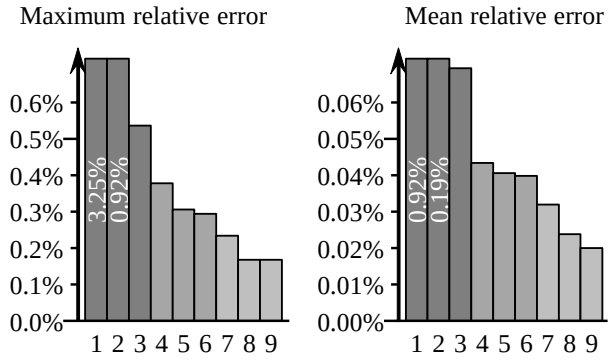

Figure 5. Relative error of the outlet temperature as a function of the number of PLS components n_c.

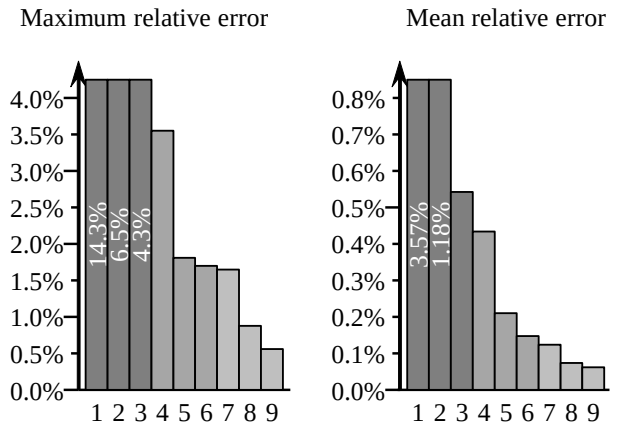

Figure 6. Relative error of the extent of reaction as a function of the number of PLS components n_c.

Unlike the outlet pressure and temperature, the extent of reaction ξ does not result in a similar good fitting as it can be seen in figure 6. This can be explained by the influence of all independent variables in the first four components defined *via* PLS indicating the difficulty to find linear combinations. This is also visible in the increase in the explained variance in the response ξ from 82.01% with $n_c = 1$ to 97.89% with $n_c = 5$. This finding correlates with the improve of the fit as it was in the case of the pressure and temperature. The maximum and mean relative error using 5 PLS components corresponds hereby to an error of 7.72 mol/s and 0.86 mol/s respectively. Despite the relatively high error in these calculations, it is possible to apply the extent of reaction surrogate model with 5 PLS components into the procedure outlined by us previously (Straus and Skogestad, 2016).

Conclusion and Outlook

The developed three-step procedure was applied to two examples, a pipe model and the reaction section of the ammonia synthesis loop. In both cases, it was possible to obtain surrogate models with high accuracy considering the reduction in the variable space. Incorporation of the surrogate model into a flowsheet consisting of a synthesis-gas make-up section, the reaction section, and a separation section results in a maximum relative error of 0.1% in all streams.

A detailed study looking into the application of this procedure to the reactor loop modelled in Aspen HYSYS is currently in development consisting of a similar flowsheet topology and 11 independent variables. The first results indicate that similar performance can be expected for this detailed model.

Acknowledgments

The authors gratefully acknowledge the financial support provided by Yara International ASA.

References

Badhurshah, R. and Samad, A. (2015). Multiple surrogate based optimization of a bidirectional impulse turbine for wave energy conversion. *Renewable Energy*, 74:749 – 760.

Biegler, L. T., Grossmann, I. E., and Westerberg, A. W. (1997). *Systematic Methods of Chemical Process Design*. Prentice Hall.

Boulesteix, A.-L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44.

Caballero, J. A. and Grossmann, I. E. (2008). An algorithm for the use of surrogate models in modular flowsheet optimization. *AIChE Journal*, 54(10):2633–2650.

Cozad, A., Sahinidis, N. V., and Miller, D. C. (2014). Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60(6):2211–2227.

Cozad, A., Sahinidis, N. V., and Miller, D. C. (2015). A combined first-principles and data-driven approach to model building. *Computers & Chemical Engineering*, 73:116 – 127.

Foss, B., Grimstad, B., and Gunnerud, V. (2015). Production optimization facilitated by divide and conquer strategies. *IFAC-PapersOnLine*, 48(6):1 – 8. 2nd {IFAC} Workshop on Automatic Control in Offshore Oil and Gas Production {OOGP} 2015Florianpolis, Brazil, 2729 May 2015.

Grimstad, B., Foss, B., Heddle, R., and Woodman, M. (2016). Global optimization of multiphase flow networks using spline surrogate models. *Computers & Chemical Engineering*, 84:237 – 254.

Straus, J. and Skogestad, S. (2016). Minimizing the complexity of surrogate models for optimization. In Kravanja, Z. and Bogataj, M., editors, *26th European Symposium on Computer Aided Process Engineering*, volume 38 of *Computer Aided Chemical Engineering*, pages 289 – 294. Elsevier.

Wold, S., Martens, H., and Wold, H. (1983). *The multivariate calibration problem in chemistry solved by the PLS method*, pages 286–293. Springer Berlin Heidelberg, Berlin, Heidelberg.