Computers and Chemical Engineering 139 (2020) 106841

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Understanding chemical production processes by using PLS path model parameters as soft sensors

Geert H. van Kollenburg[a,b,*], Jacoline van Es[a], Jan Gerretzen[c], Heleen Lanters[a,b], Roel Bouman[a,b], Willem Koelewijn[c], Anthony N. Davies[c,d], Lutgarde M.C. Buydens[a], Henk-Jan van Manen[a,c], Jeroen J. Jansen[a]

[a] Radboud University, Department of Analytical Chemistry/ Chemometrics, Institute for Molecules and Materials (IMM), Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
[b] TI-COAST, science park 904, 1098 XH Amsterdam, The Netherlands
[c] Nouryon Chemicals B.V., Supply Chain, Research & Development, Expert Capability Group Measurement & Analytical Science, Zutphenseweg 10, 7418 AJ Deventer, the Netherlands
[d] University of South Wales, Faculty of Computing, Engineering and Science, CF37 1DL Pontypridd

ARTICLE INFO

Article history:
Received 26 November 2019
Revised 20 February 2020
Accepted 29 March 2020
Available online 1 May 2020

Keywords:
PAT
Soft sensors
PLS-PM
Predictive modelling
Process analytics
Chemometrics

ABSTRACT

To make industrial processes lean, inclusion of technical process information is required into statistical modelling. Understanding how parts of a process are related to other parts and to output quality is key to understanding and controlling processes. In this work, we show how PLS path modelling can be used to incorporate process knowledge into predictive chemical process analysis. The result is a wealth of information which is not obtained by standard data analytic techniques commonly used by analytical chemists or process engineers. By comparing model parameters across multiple data sets from different batches of the same process, model parameters could be used as soft sensors. Some variables which would normally be discarded as uninformative were highly predictive of production costs. The methodology reported here improves chemical process understanding through the analysis of complementary historical process data, which may serve as the basis for development of improved process conditions and control.

1. Introduction

There is a consistent societal and political push for industrial processes to reduce resource consumption. This reduction needs to be attained within economically viable settings for the producers. One of the main principles of the current transition to Industry 4.0 is to make processes lean, meaning reduced waste and consistent end-product quality (Hermann et al., 2016). Understanding the effects of the raw material variations, the production settings and process conditions on the quality of end product is key to comprehensive process understanding and control (Qin, 2012). Any information source that may contribute to smart process operation is therefore highly valuable.

Partial Least Squares (PLS) Regression (Geladi and Kowalski, 1986; Wold et al., 1983; Wold et al., 1984; Wold et al., 2001) is one of the standard tools in the arsenal of analytical chemists to make predictions based on the large amounts of data that are collected during production processes. The PLS method can extract information from a wide variety of variables (e.g., spectral and/or high dimensional data). Its flexibility and implementation in many standard software packages have made this method the standard for predictions in multivariate analytical data. By extracting the most predictive, correlative information out of a large number of variables, PLS regression is powerful in monitoring and prediction contexts. However, the downside is that the model treats all measurements without hierarchy or conditionality. The PLS model therefore provides limited understanding about the process that generates the variability in the measured data and about the effects that each variable has on the process (Wold et al., 1983). A wealth of knowledge therefore remains untouched because the information on technical aspects of processes is not incorporated into the predictive modelling.

Production processes that comprise multiple process units may enable separation of the process variables into unit-specific blocks. Multi-block (MacGregor et al., 1994; Westerhuis et al., 1998) and

* Corresponding author at: Radboud University, Department of Analytical Chemistry/ Chemometrics, Institute for Molecules and Materials (IMM), Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands.
E-mail address: g.vankollenburg@science.ru.nl (G.H. van Kollenburg).

https://doi.org/10.1016/j.compchemeng.2020.106841

orthogonalised PLS methods (Menichelli et al., 2014; Næs et al., 2011; Romano et al., 2019) exist that can be used to evaluate the effects of each block on the output. These methods impose a hierarchy on the blocks, such that specific blocks are always used first in the predictions. The most prominent drawback of these methods in process analysis is that the predictor blocks are, by definition, independent. The various units of a production process are often highly interrelated (e.g., temperatures of subsequent blocks will be highly related) and therefore the statistical modelling should take those relationships into account.

To accommodate the interrelated predictor blocks of a chemical production process, application of PLS-Path Modelling (PLS-PM) (Wold, 1982) for the statistical analysis of such production processes is proposed here. PLS-PM is a general soft-modelling approach which originated in the social sciences and is perfectly suited for modelling multi-block processes where the directionality of associations is clear. Therefore, path models are powerful tools in understanding the effects of multi-block processes (Kowalski et al., 1982; Tenenhaus & Hanafi, 2010). While other multi-block PLS models have been proposed repeatedly to analyse chemical processes (Næs et al., 2011; Qin, 2012; Qin et al., 2001; Wangen and Kowalski, 1989; Westerhuis et al., 1998), PLS-PM itself has seen little to no use in understanding in detail the relationships within a process. As will be shown, PLS-PM can benefit our understanding of the inner workings of a chemical production process by incorporating process knowledge into the predictive modelling. PLS-PM may be used as an addition to graph-based models, such as signed directed graphs (Chiang et al., 2000; Kramer & Palowitch Jr, 1987) and may help in developing models for fault diagnosis (Lucke et al., 2020; Russell et al., 2000).

This paper considers the analysis of a semi-batch chemical production process at Nouryon (previously AkzoNobel Specialty Chemicals). The process comprises a number of interconnected units, like heaters, coolers, and reaction vessels. The production process runs until the catalytic efficiency is below a certain threshold. The process is then stopped, a new batch of catalyst is introduced and the production resumes. The yield per batch of catalyst, and accordingly the production cost, is highly variable.

The question that strongly drove this research was whether model parameters could be found which were related to resource consumption. In other words, is it possible to use the model parameters themselves as an indication that the costs will be high or low? Fitting the same PLS-PM model to process data from multiple batches enabled us to increase our understanding of how the units in the production process are interrelated, improving the possibilities for process control, and at the same time model parameters could be related to variations in production costs. In this sense, this research provided a method to use model parameters (rather than process variables) as soft sensors, something which to our knowledge has not been reported before.

The power of using combination of variables as soft sensors in production processes is thoroughly established. Using model parameters as soft-sensors may provide much more information about the actions to take when something goes wrong. Since the model provides information about the relations of one part of the process to the other parts, this will enable process operators to better substantiate decisions about how to control the process.

2. The PLS path model

The first step of analysing a process with PLS-PM is to partition the manifest (i.e. measured) variables (MVs) into Q blocks. This partitioning of variables is called the measurement model (or outer model). Each block of MVs is summarized by a single latent variable (LV), ξ_q. The blocks (and thus the LVs) are connected to other blocks through the structural (inner) model in terms of regression equations. Some blocks are only used as predictors (the exogenous blocks) and some blocks are used as responses (the endogenous blocks). The structural model is estimated as a series of ordinary linear regression models, one for each endogenous block. For the regression of an endogenous block summarized by the LV ξ_q the LVs belonging to its predictors are collected in the matrix $\boldsymbol{\Xi}_q^*$. The prediction of $\xi_{\boldsymbol{q}}$ can be written as

$$\hat{\xi}_{\mathrm{q}} = \boldsymbol{\Xi}_q^* \boldsymbol{b}_q$$

where the vector of regression effects, $\boldsymbol{b}_g$, is found through ordinary least squares as

$$\boldsymbol{b}_g = \left(\boldsymbol{\Xi}_q^{*T}\boldsymbol{\Xi}_{\boldsymbol{q}}\right)^{-1}\ \boldsymbol{\Xi}_{\boldsymbol{q}}^{*T}\xi_q.$$

In PLS-PM, LVs are estimated through an iterative PLS procedure which optimally reproduces the covariance between all manifest variables. In contrast to the often used sequential-and-orthogonalised(SO) - PLS methodology, there is no hierarchical ordering in the importance of each block. The PLS-PM algorithm as described here is implemented in the "plspm" package by Sanchez (2013) and Sanchez et al. (2017) for version 3.5.1 of the R programming language (R Core Team, 2018). This software is able to natively handle missing values and non-normal variables, which are omnipresent in process analytical data.

There are a number of options to choose from when estimating a PLS-PM. In the current application, the measurement model used the reflective mode (Vinzi et al., 2010). For the structural model the centroid scheme was used to estimate latent variable as it does not overestimate effects, has good convergence and performs well for large sample sizes (Wilson and Henseler, 2007). For more technical details on the estimation algorithm and model options the interested reader is referred to other works (Tenenhaus and Vinzi, 2005; Vinzi et al., 2010).

3. Implementation of PLS-PM into predictive modelling

3.1. Data

The data under analysis originates from a semi-batch production process at Nouryon. More than 21,000 hourly measurements of process variables (e.g. temperatures, flow rates, pressures) were collected at each of the units in the production process. This data is complementary to the catalyst performance and were originally collected for other purposes. Additionally, data on end-product quality was also available. These data were not measured hourly, but measured with a variable frequency (details follow in Section 3.3).

The steps of the process (cooling, heating and catalysed reaction), take place in physically different parts of the factory. The chemical mixture is transported from one part to the next through a series of pipes. The reaction involves a heterogeneous catalyst which degrades over time. Currently, catalyst performance is monitored according to a particular pressure drop. If this pressure drop is outside predefined limit values, the process is stopped, the catalyst is replaced with a new catalyst batch and production resumes. The lifetime of this catalyst is related to production cost and experience shows that this lifetime is highly variable between batches. The dataset was separated according to the corresponding batch of catalyst. The term 'batch' is therefore also used for a subset of data related to a batch of catalyst.

A total of eleven batches were analysed. For the current application, production costs are defined as how much product can be made with one batch of catalyst. The production costs of the batches varied from 37 to 208 (in arbitrary units); lower values meaning lower costs (see Table 1). Note that if the catalyst can be used a long time before having to be replaced, the costs are lower

Table 1
Number of hourly measurements per batch.

Batch no.	Production cost	Sample size
1 ('Best')	37	3402
2	38	3610
3	39	3349
4	45	3087
5	57	2100
6	64	1750
7	100	1278
8	133	1064
9	138	873
10	182	580
11 ('Worst')	208	581

and at the same time there will be more hourly measurements for that batch. This inverse relationship between cost and number of hourly measurements is not exact because the production rate is not always constant and because the number of measurements that had to be removed in the preprocessing steps varied per batch (see Section 3.2).

3.2. Data preprocessing

In order to effectively compare the PLS-PM results between batches, process variables were removed if they had zero variance in one or more batches, or when they had such low variance that the resulting models could not be statistically validated. A total of 37 process variables were considered in the final analyses. Hourly measurements that were collected while the plant was not in operation were also removed. Sample size numbers for the batches were determined by batch length, and ranged from 580 to 3610 hourly measurements (see Table 1).

3.3. Pathway determination

The production process that was analysed comprises six production units such as heaters, coolers and tanks. Additionally, a flow control unit exists which regulates the process. At each of these seven blocks a number of process variables are continuously measured (see below). In addition to these seven process blocks, the first block in the model comprised a set of variables related to the characteristics of the input material—these are also continuously monitored and available as hourly data. As the last block in the model, the set of variables related to the product quality was incorporated. Fig. 1 shows the fully specified model with manifest variables attached to each block. The actual process names of the variables cannot be disclosed, so in the results section we will number them V1 to V37 in the order they appear in this model, starting with the variables related to 'Input' and ending with the product quality variables at block 'Quality'.

To avoid confusion, note that the variables at block 'Reaction' are not variables from which catalyst performance was calculated. These variables are the process measurements which were constantly measured at the unit in which the catalysed reaction takes place (i.e., pressure, temperature, flowrates). In the same light, the variables at the 'Quality' block are quality measurements like purity of the product, and do not include the production costs (the 'Cost' variable only has a single value per batch and will be used later to relate model parameters to).

In Fig. 1, the inner model contains many arrows which indicate direct effects of one block on another. We formulated this inner model by taking into account three criteria (see Fig. 2). Firstly, the direct effects were specified that were associated to the physical architecture of the process (i.e., the piping). Secondly, theoretical considerations and expert knowledge from the process engineers were used to define additional paths between the blocks. Thirdly, paths were included to predict the direct effects of every process block on the end-product quality.

A note on the Quality block is required here as the variables of the Quality block were not measured hourly, but on a less frequent basis. In practice, the last measured values of the quality variables are used as a best guess for their current value, until the values are updated by making new measurements. As we are using historical data, we were able to incorporate this practice into the modelling procedure as follows. Since the process variables are related to the quality of the product by design, we used the hourly measurements to predict what the values of the Quality variables would be at the each next update. Because the measurement frequency is not the same for the hourly process variables and the Quality variables, the actual regression relates the medians of each process variables between updates to the current values of the Quality variables. While this may provide a robust prediction from the

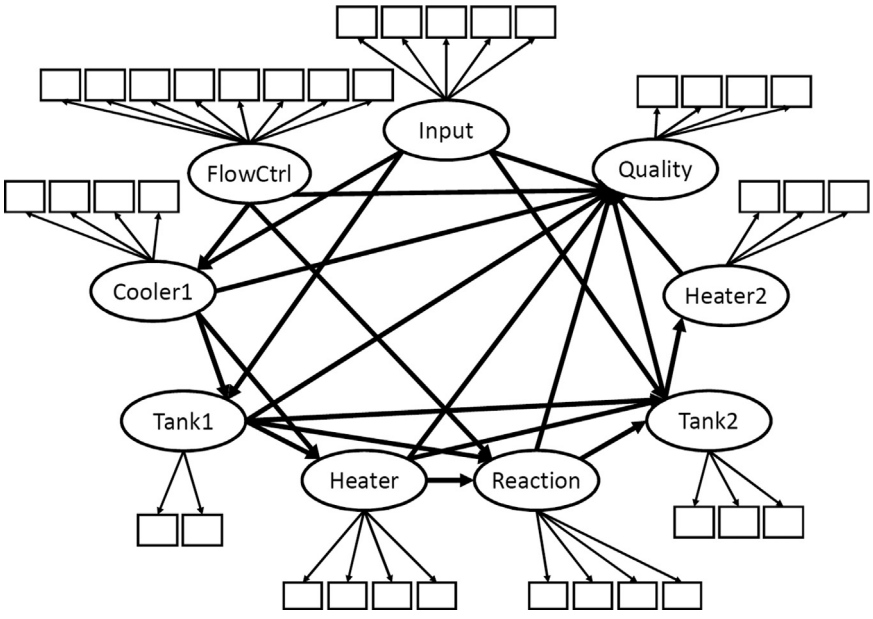

Fig. 1. Full PLS-PM used in this research. Squares represent the manifest process variables; ellipses represent latent variables.

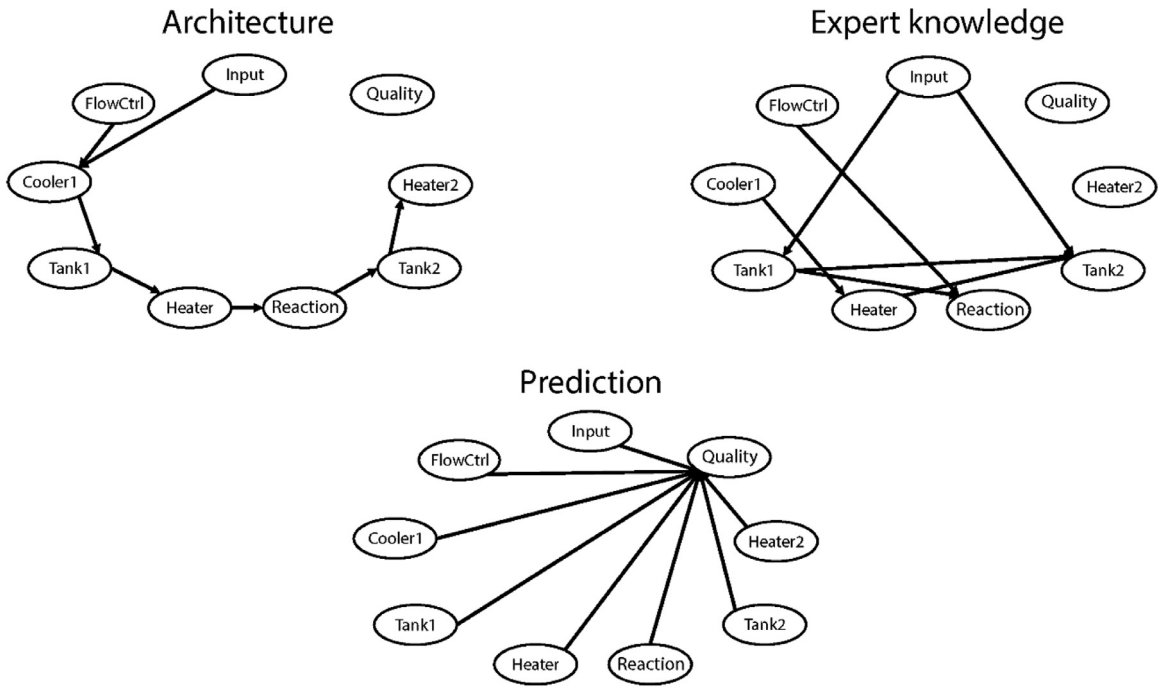

Fig. 2. Three criteria for inclusion of effects in the inner model of the PLS-PM. The arrows indicate direct regression effects.

process measurements, the standard errors for the regression coefficients will be underestimated and RMSE scores will also not be valid. These drawbacks only apply to the regression of the quality parameters.

3.4. Validation of the structural model

The model described in Section 3.3 was estimated for each of eleven batches separately. Due to the strong variation in production costs it was expected to find large differences in estimated model parameters, *i.e.*, in the estimated relationships between the blocks of the process. Of course, many other model specifications may be made with other considerations and those models could be compared to see what model fits best to the data. However, the current application was confirmative in nature, meaning that interest was not in finding new connections, but to evaluate whether the current theoretical considerations are in agreement with the data.

Unfortunately, there is no straightforward method to test model fit in PLS-PM (Henseler and Sarstedt, 2013). The paths in the structural model were therefore evaluated by checking whether they could explain significant variance in their corresponding endogenous block. Since multiple batches of data were analysed, paths that do not contribute to explained variance in any batch only lead to unnecessary model complexity. None of the pre-defined paths shown in Fig. 2 were non-significant in all the batches so they were all kept in the model.

3.5. Variable importance

Additionally, relationships between individual manifest variables and the quality variables were tested. Note again that the quality variables do not represent production costs or catalyst performance, but rather the chemical quality such as purity of the product. There are many indirect pathways going from one manifest variable to another. From a path model, it is possible to calculate the (reproduced) association between individual variables using standard rules for combining indirect effects (Keith, 2014). Again, these associations can be used to pinpoint process faults on a detailed level, but it is beyond the scope of the presented research to provide details on what specific variables indicate.

3.6. Comparison to PLS(2) regression

The method of using model parameters as soft-sensors can be applied to virtually any statistical model. For comparison purposes, the data was analysed by PLS regression analysis (Wold et al., 2001). The same 37 process variables were used in the predictor block, and the four product quality parameters were used in the response block (hence, PLS2 was used). The variables were autoscaled before analysis. The maximum number of LVs for the predictors was set to 7 (i.e., equal to the number of blocks and LVs in PLS-PM). Since there is no information in PLS regression which compares well to the inner model coefficients of PLS-PM, we opted for a measure of variable importance, which could be related to the variable importance described in Section 3.5.

Assessment of influential variables is common practice in PLS regression. We used the Variable Importance in Projection (VIP) score (Eriksson et al., 2013), which is calculated for each variable x_j as:

$$VIP_j = \sqrt{\frac{J}{\sum_{m=1}^{M} R^2_{Y.\boldsymbol{z_m}}} \times \left(\sum_{m=1}^{M} w^2_{jm} \times R^2_{Y.\boldsymbol{z_m}} \right)},$$

where J is the number of variables, $R^2_{Y.z_m}$ is the amount of Y variance explained by the m-th latent variable, and w^2_{jm} is the squared weight (importance) of variable j on the m-th latent variable. If a process variable is related to high production costs, one would expect a positive correlation between a VIP score and cost.

3.7. Explained variance in product quality

An important aspect of a PLS model (and other regression models) is to evaluate how well the predictor variables can explain the response variable. Next to the individual effects of the predictors

Table 2
Average regression effects across batches and the correlation between the effects and cost.

Response block	Predictor block	Mean (sd) of regression coefficient	Correlation between coefficient and cost.
Cooler1	Input	0.01 (0.26)	0.50
	FlowControl	0.30 (0.68)	0.42
Tank1	Input	0.13 (0.25)	−0.40
	Cooler1	−0.21 (0.64)	0.55
Heater	Cooler1	−0.49 (0.59)	−0.56
	Tank1	0.03 (0.35)	−0.39
Reaction	FlowControl	0.12 (0.29)	−0.18
	Tank1	−0.22 (0.29)	−0.11
	Heater	−0.07 (0.59)	0.07
Tank2	Input	−0.01 (0.20)	−0.22
	Tank1	−0.05 (0.46)	0.08
	Heater	−0.07 (0.39)	0.83
	Reaction	0.21 (0.24)	0.42
Heater2	Tank2	−0.68 (0.53)	−0.15
Quality	Input	0.10 (0.09)	0.28
	FlowControl	0.19(0.15)	0.18
	Cooler1	0.39 (0.29)	0.21
	Tank1	0.18 (0.12)	0.06
	Heater	0.33 (0.19)	0.70
	Raction	0.22 (0.19)	0.06
	Tank2	0.39 (0.26)	0.05
	Heater2	0.27 (0.28)	0.18

on each outcome variable, it was evaluated how well the quality variables could be predicted from the process variables. In PLS2 regression, all variables were combined together to explain the variance in the quality block. In PLS-PM the variables were first combined according to the block structure, and the quality block was regressed on the LVs of all the other blocks (cf. Fig. 2, Prediction). Both PLS2 regression and PLS-PM provided explained variance as standard analysis output.

4. Results

4.1. Structural model

Table 2 shows the results of relating the inner model coefficients of PLS-PM to production cost. To understand the table better, let us review how PLS-PM works. Each block of MVs is summarised in a single LV. The structural model is created by regressing the LV of a response block (Column 1 in the Table 2) on the LVs of the corresponding predictor blocks (Column 2). Column 3 of Table 2 indicates the average regression coefficient over the 11 batches. Many regression coefficients had substantial variation across the batches. The standard deviation of the regression is reported in parentheses in Column 3. The variation in the regression coefficient across the batches were related to the variation in production costs of the batches. The correlation between the regression coefficients and the production costs (there was a single cost value per batch) is found in Column 4.

The use of p-values is hardly meaningful here. This research is based on the available historical data and there are a limited number of 11 batches. The use of significance levels (and cut-offs) may lead to incorrect inclusion or exclusion of associations for possible follow-up research. Therefore, in line with the view of the American Statistical Association, p-values are not reported for these correlations (Wasserstein and Lazar, 2016).

Table 2 shows that the strongest correlations with production costs were found in the effects of Heater on Tank2 and on Quality. The effect of Heater on Tank2 in the different batches was on average close to zero, but an increase in this effect related to higher production costs. Heater had on average a moderate effect on the Quality. When this effect increased, this was related to higher costs as well. Vice versa, in batches with a lower cost, the effects were also lower. These effects are very important from a process control viewpoint, since the heating is such a crucial part of the production process.

Table 3
Notable relations between importance of manifest variables in PLS-PM and production costs.

Variable	Mean(sd) loading	Correlation between loading and cost
V7	0.26 (0.24)	0.52
V10	0.86 (0.19)	−0.46
V12	0.88 (0.20)	−0.46
V14	0.45 (0.26)	0.57
V24	0.57 (0.28)	0.47
V25	0.71 (0.30)	0.43
V26	0.71 (0.25)	−0.49
V27	0.68 (0.25)	0.43
V29	0.16 (0.14)	0.78
V31	0.87 (0.07)	−0.49
V35	0.70 (0.15)	0.43
V37	0.64 (0.20)	0.48

It is interesting to note that relations with Cooler1, which is at the beginning of the process, is strongly correlated with the catalytic efficiency, and thereby with production costs. This indicates that monitoring and control of the first parts of the process may be extremely important in keeping the catalyst working optimally.

4.2. Measurement model

In the measurement model, a loading (or, weight) is calculated for each MV which indicates how important the MV is for the corresponding LV. In Table 3 we show the average loadings (Column 2) for a number of variables (Column 1). Like before, the loadings varied across the 11 batches (see parentheses in column 2 for the standard deviation). For some variables, the variation in loadings was correlated with the costs (Column 3). Table 3 provides the variables for which this correlation was greater than 0.4. Some variables related to costs have consistently high loadings (i.e., V10, V12 and V31), while others have much more variable loadings (i.e., V14, V24 and V22).

Importantly, from a soft-sensor point of view, variables like V7 and V29 are of tremendous interest. On average these variables

Table 4
Notable relations between VIPs and production costs.

Variable	Mean(sd) of VIP	Correlation VIP and Cost
V1	1.03 (0.27)	−0.52
V3	0.90 (0.26)	−0.50
V25	1.09 (0.25)	−0.51
V31	0.97 (0.18)	0.57
V32	1.03 (0.47)	0.55

contribute very little to the model (i.e., their loading is generally low). Standard methodology would likely regard these variables as uninformative. But when we relate the loadings to production costs, we see that these variables are strongly related to those costs. A straightforward explanation is that in costly batches these variables show greater variance than in low cost batches. So instead of disregarding these variables because their contribution to the model is limited, we should instead focus extra on controlling these variables.

4.3. PLS2 regression

The parameters from the PLS2 regression, which excludes block information and uses all process measurements as a single input matrix, were also used as soft-sensors. The contribution of each variable to the model was calculated by the VIP and this contribution was then related to production costs. Five variables were found for which this correlation was greater than 0.4 (see Table 4). The spread of the VIPs was rather large and none of the variables had a consistently strong contribution (a VIP larger than one indicates important contribution). Variables 25 and 31 show up in both methodologies (cf. Tables 3 and 4) as being related to production costs, yet with opposite signs. We have yet to understand whether this connection can be explained in terms of the algorithms of the modelling procedures.

4.4. Explained variance

PLS-PM combines multiple regression models, while PLS regression only regresses a single Quality block onto all process variables simultaneously. The only direct comparison between the two is therefore the variance in the Quality block that can be explained by the model. That is, how well can the process variables predict new product quality measurements? Table 5 shows how much variance of the quality block each model could explain in the various batches. In eight out of the eleven batches, PLS-PM was able to explain more variance of the quality block.

It was also checked whether the explained variance of the Quality block was related to production cost. For PLS PM there was no apparent relation between explained variance and cost. For PLS2 regression there was a correlation of approximately 0.5 between explained variance and costs (Fig. 3). Two clear deviations from the linear relationship for could be seen for Batch 3 and 6. When these batches are not considered, the correlation between cost and explained variance by the PLS2 model becomes very strong (i.e., 0.862, See Fig. 4). Investigations are underway to evaluate Batches 3 and 6 in detail.

Table 5
Explained variance of the product by PLS-PM and PLS regression. Batches are ordered from low to high cost.

		$R^2_{quality}$	
Batch no.	*Cost*	PLS PM	PLS2
1	37	0.24	0.21
2	38	0.10	0.20
3	39	0.50	0.37
4	45	0.18	0.10
5	57	0.46	0.12
6	64	0.68	0.42
7	100	0.16	0.28
8	133	0.47	0.24
9	138	0.75	0.25
10	182	0.54	0.42
11	208	0.23	0.38
Correlation R^2 and cost		0.221	0.499

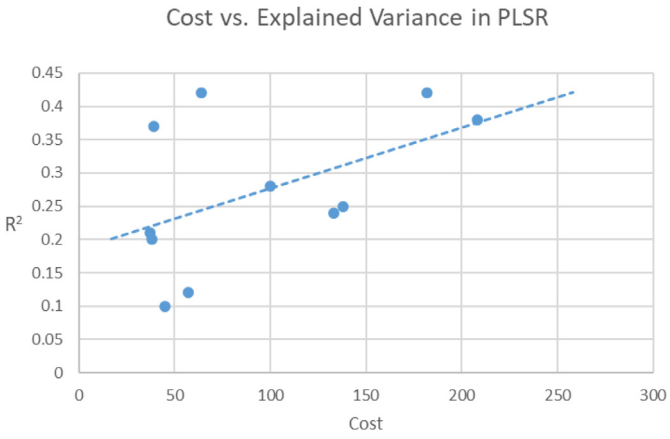

Fig. 3. Cost versus explained variance in PLS2.

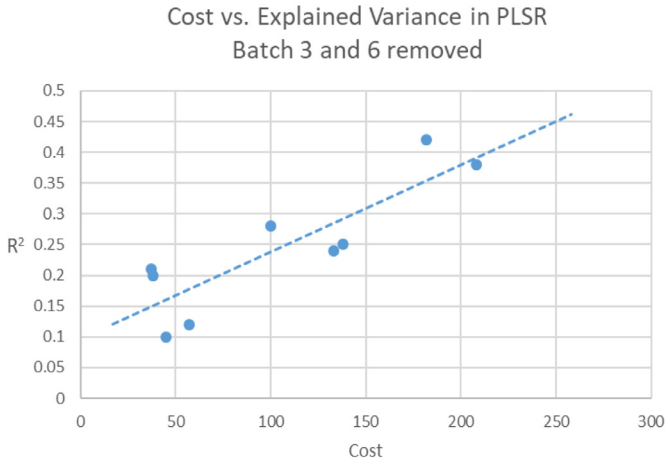

Fig. 4. Cost versus explained variance in PLS2 with two batches discarded.

5. Discussion

One of the key aspects of PLS-PM is that only a single latent variable is considered per block. Though this method inherently protects us from overfitting (i.e., the total number of latent variables is always equal to the number of specified blocks), it can be an issue when the variables within a block are not, or hardly correlated or when variables of different blocks are correlated (leading to high cross-loadings). We expect that the performance of the model to accurately describe the important factors in the production factory will improve when we incorporate multiple latent variables per block. Future applications of path modelling approaches for process analytics should focus on both the interrelations between blocks (as PLS-PM does) as well as the possibility to obtain multiple LVs per block (as in most other multi-block PLS methods). Work is being done by the current authors to develop Process PLS, a PLS-based path model which is suited for multi-dimensional blocks, flexible model specifications, and exploratory

and confirmatory applications (van Kollenburg et al., In preparation).

An additional value of a path modelling approach within process analysis is that such analyses provide information on indirect effects. With PLS PM (and with other graph-based methods) these indirect effects can be used to predict how fluctuations early in the process may affect the rest of the process. Additionally, the model predictions and the information resulting from using model parameters as soft-sensors could be tested in pilot plants (this holds for applications of virtually any model estimated on multiple batches). Such experiments may provide insight in causal pathways that were not considered from a theoretical point of view before. Future developments may also consider applications for streaming data (where data points come in on-the-fly). By continuously updating and monitoring model parameters, the applicability of process analytical models will be increased even more.

We also evaluated whether PLS2 regression VIP scores could be used as soft-sensors, but there were no strong relationships between the VIP scores and costs. The relative difficulty to find the important variables with PLS2 regression shows how much can be gained in process analysis when researchers include process information to separate variables and estimate a PLS PM, rather than combining all variables and interpreting PLS2 regression results as is done in standard practice.

Some of the presented results showed that explained variance of product quality was related to the costs associated to the batches. This can to a certain extend be explained by the fact that the goal of a production process is generally to produce a product of constant quality. In a (theoretical) perfect manufacturing process, fluctuations in process variable only exist to ensure that product quality is constant. If in such cased, quality does not have any variance to explain and any correlation between process variables and product quality will be undefined (or 0, depending on the definition used). Any effect of process variables on the product quality will move away from the no-correlation situation. We saw similar trends in our results. If we directly predicted product quality from all the process variables (using PLS2), more variance can be explained in worse batches, meaning that some (combinations of) process variables affect the product quality. This trend is less clear in the path modelling approach, because product quality in PLS PM was only one of the dependent blocks that needed to be explained. Additionally, the relationships between the other process blocks are specifically not independent, since all process variables work together to obtain the desired product quality.

On a final note, the regression part of the path modelling approach was justified in part by the assumption that causality in the production process goes one way. However, variables at a later stage of the process may let a process controller decide to change settings from an earlier part of the process. For example, a pressure which becomes too high in the middle of the process may influence what setting is chosen at the flow control. Such indirect effects caused by manual adaptations are difficult to incorporate into the statistical model even if each decision would have been well-documented. Process engineers are not required to document each decision in their daily routine, which makes reconstruction of these feedback loops challenging. Experience has shown that different process operators tend to have different techniques to control the process. A suggestion for future research is to incorporate these feedback loops by using, for instance, a mixture of non-recursive regression models.

6. Conclusions

In this paper PLS-PM was used to model an industrial chemical production process of Nouryon. PLS-PM enables researchers to separate blocks of variables according to where in a production process they were measured. In this way one can obtain detailed information on the interrelationships between the various parts of the process. By analysing eleven batches of data and using the model parameters as soft-sensors, it was shown that specific variables and connections between production units were highly related to production costs.

PLS-PM was chosen over other PLS regression techniques because the former can provide detailed insights into the relationships between parts of a production process. Additionally, we were able to evaluate differences in these relationships across different production batches. These differences could then be related to the production costs associated with each batch, leading to a more detailed level of process understanding.

PLS-PM was able to show which relationships between particular blocks should be controlled very firmly, as they are related strongly to the production cost. We also showed that by relating variable contributions to production costs (or other external variables) can identify important variables which otherwise would be deemed uninformative.

We showed that PLS-PM was, overall, better able to explain the variation in product quality (as evaluated through explained variance) than PLS regression. Taking into account the wealth of additional information that can be obtained in PLS-PM that cannot be obtained from PLS regression (i.e., everything related to the inner model of PLS-PM), we are confident that incorporating process knowledge into the PLS framework may lead to better process understanding and more powerful statistical models in the future.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Geert H. van Kollenburg: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing - original draft. **Jacoline van Es:** Investigation, Formal analysis, Validation, Data curation, Software. **Jan Gerretzen:** Data curation, Resources, Project administration, Writing - review & editing. **Heleen Lanters:** Investigation, Formal analysis, Data curation, Methodology, Software. **Roel Bouman:** Investigation, Validation, Visualization, Writing - review & editing. **Willem Koelewijn:** Investigation, Resources. **Anthony N. Davies:** Funding acquisition, Resources, Supervision. **Lutgarde M.C. Buydens:** Funding acquisition. **Henk-Jan van Manen:** Conceptualization, Funding acquisition, Resources, Supervision, Writing - review & editing. **Jeroen J. Jansen:** Funding acquisition, Project administration, Supervision, Writing - review & editing.

Acknowledgments

This research was in part funded by the Netherlands Organization for Scientific Research (NWO) through the PTA-COAST3 “Outfitting the Factory of the Future with Online analysis” (OFF/On) consortium. The authors would like to thank Kamiel Mellema and Rianne Timmermans for their shared expertise on the modelled production process.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2020.106841.

References

Chiang, L.H., Russell, E.L., Braatz, R.D., 2000. Fault Detection and Diagnosis in Industrial Systems. Springer Science & Business Media.

Eriksson, L., Byrne, T., Johansson, E., Trygg, J., Vikström, C., 2013. Multi- and Megavariate Data Analysis Basic Principles and Applications. Umeå: Umetrics Academy.

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 185, 1–17.

Henseler, J., Sarstedt, M., 2013. Goodness-of-fit indices for partial least squares path modeling. Comput. Stat. 28, 565–580.

Hermann, M., Pentek, T., Otto, B., 2016. Design Principles for Industrie 4.0 Scenarios. In: 49th Hawaii International Conference on System Sciences (HICSS). IEEE, pp. 3928–3937.

Keith, T.Z., 2014. Multiple Regression and Beyond: An Introduction to Multiple Regression and Structural Equation Modeling, (2nd ed.) Routledge, New York.

Kowalski, B.R., Gerlach, R.W., Wold, H., 1982. Chemical systems under indirect observation. In: Jöreskog, K.G., Wold, H. (Eds.), Systems under indirect observation: Causality - Structure - Prediction, Part II. Amsterdam: North-Holland, pp. 191–207.

Kramer, M.A., Palowitch Jr, B., 1987. A rule-based approach to fault diagnosis using the signed directed graph. AIChE J. 33, 1067–1078.

Lucke, M., Stief, A., Chioua, M., Ottewill, J.R., Thornhill, N.F., 2020. Fault detection and identification combining process measurements and statistical alarms. Control Eng. Pract. 94, 104195.

MacGregor, J.F., Jaeckle, C., Kiparissides, C., Koutoudi, M., 1994. Process monitoring and diagnosis by multiblock PLS methods. AIChE J. 40, 826–838.

Menichelli, E., Almøy, T., Tomic, O., Olsen, N.V., Næs, T., 2014. SO-PLS as an exploratory tool for path modelling. Food Qual. Preference 36, 122–134.

Næs, T., Tomic, O., Mevik, B.H., Martens, H., 2011. Path modelling by sequential PLS regression. J. Chemom. 25, 28–40.

Qin, S.J., 2012. Survey on data-driven industrial process monitoring and diagnosis. Ann. Rev. Control 36, 220–234.

Qin, S.J., Valle, S., Piovoso, M.J., 2001. On unifying multiblock analysis with application to decentralized process monitoring. J. Chemom. 15, 715–742.

Core Team, R, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Romano, R., Tomic, O., Liland, K.H., Smilde, A., Næs, T., 2019. A comparison of two PLS-based approaches to structural equation modeling. J. Chemom. e3105.

Russell, E.L., Chiang, L.H., Braatz, R.D., 2000. Data-Driven Methods for Fault Detection and Diagnosis in Chemical Processes. Springer Verlag, London.

Sanchez, G., 2013. PLS Path Modeling with R. Trowchez Editions, Berkeley, p. 383.

Sanchez, G., Trinchera, L., & Russolillo, G. (2017). plspm: tools for Partial Least Squares Path Modeling (PLS-PM). R package version 0.4.9.

Tenenhaus, M., Hanafi, M., 2010. A bridge between PLS path modeling and multi-block data analysis. In: Handbook of Partial Least Squares. Springer, Berlin, pp. 99–123.

Tenenhaus, M., Vinzi, V.E., 2005. PLS regression, PLS path modeling and generalized Procrustean analysis: a combined approach for multiblock analysis. J. Chemom. 19, 145–153.

van Kollenburg, G. H., Bouman, R., Gerretzen, J., van Manen, H., & Jansen, J. J. (In preparation). Process PLS: Incorporating Process Knowledge in the Analysis of Multiblock, Multidimensional and Multicollinear Data.

Vinzi, V.E., Trinchera, L., Amato, S., 2010. PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement. In: Handbook of Partial Least Squares. Springer, Berlin, pp. 47–82.

Wangen, L., Kowalski, B., 1989. A multiblock partial least squares algorithm for investigating complex chemical systems. J. Chemom. 3, 3–20.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p-values: context, process, and purpose. Amer. Stat. 70, 129–133.

Westerhuis, J.A., Kourti, T., MacGregor, J.F., 1998. Analysis of multiblock and hierarchical PCA and PLS models. J. Chemom. 12, 301–321.

Wilson, B., Henseler, J., 2007. Modeling reflective higher-order constructs using three approaches with PLS path modeling: a Monte Carlo comparison. In: ANZMAC 2007: Conference Proceedings and Refereed Papers. ANZMAC, Dunedin, pp. 791–800.

Wold, H., 1982. Soft modeling: the basic design and some extensions. In: Jöreskog, K.G., Wold, H. (Eds.), Systems Under Indirect Observation: Causality - Structure - Prediction, Part II. Amsterdam: North-Holland, pp. 1–54.

Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration problem in chemistry solved by the PLS method. In: Kågström, B., Ruhe, A. (Eds.), Matrix Pencils. Springer, Berlin, pp. 286–293.

Wold, S., Ruhe, A., Wold, H., Dunn WJ., I., 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 5, 735–743.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130.