Metin Türkay, Erdal Aydin (Eds.)
Proceedings of the 31th European Symposium on Computer Aided Process Engineering
June 6th to 9th, 2021, Istanbul, Turkey.

Data-based Industrial Soft-sensor Design via Optimal Subset Selection

Martin Mojto[a,*], Karol L'ubušký[b], Miroslav Fikar[a] and Radoslav Paulen[a]

[a]*Slovak University of Technology in Bratislava, Bratislava 81237, Slovakia*
[b]*Slovnaft, a.s., Bratislava 82412, Slovakia*
martin.mojto@stuba.sk

Abstract

Inferential (or soft) sensors are used in industry to infer the values of imprecisely and rarely measured (or completely unmeasured) variables from variables measured online (e.g., pressures, temperatures). The main challenge, akin to classical model overfitting, in designing an effective inferential sensor is to select a correct structure represented by the number of sensor inputs. This work is focused on the design of an inferential sensor for bottom product composition of an industrial distillation column. We study effectiveness of various subset selection methods that regard different model-overfitting criteria. Our results show that the subset selection is a viable methodology to sensor design and that we are able to improve accuracy of the current refinery sensor by around 15 %.

Keywords: Inferential (Soft) Sensors, Process Monitoring, Subset Selection

1. Introduction

The accuracy and reliability of industrial measurements have a huge impact on the effectiveness of industrial process control (Khatibisepehr et al., 2013). Especially, the control performance of advanced process controllers (Qin and Badgwell, 2003) is highly related to the indication quality of controlled variables (CVs). It is often the case that the crucial CVs (e.g. distillate purity) are too expensive or impossible to measure at the frequency required for an effective feedback control. This gave rise to a use of so-called inferential (or soft) sensors (Mejdell and Skogestad, 1991; Kordon et al., 2003; Curreri et al., 2020).

The purpose of an inferential sensor is to infer the CV value (output) using the data from other measured variables (inputs). The design procedure aims at a) identifying a sensor structure and b) at estimating the sensor parameters. While the latter problem can be solved relatively easily, the former issue of structure selection can be much more challenging in practice.

The focus of the paper is on a class of subset selection (SS) methods. These methods use mixed-integer programming to determine the optimal structure of an inferential sensor using various model-overfitting criteria such as adjusted R^2 (R^2_{adj}), corrected Akaike information criterion (AIC_C), Bayesian information criterion (BIC), and cross-validation. We make a comparison of effectiveness of the SS methods investigating a linear soft-

Acknowledgments: This research is funded by the Slovak Research and Development Agency under the project APVV SK-FR-2019-0004 and by the Scientific Grant Agency of the Slovak Republic under the grant 1/0691/21.

sensor design for a depropanizer column in an industrial fluid catalytic cracking (FCC) unit.

2. Problem Description

Our goal is to identify models of inferential sensors of the following linear form:

$$y = m(a_1, a_2, \ldots, a_{n_p})^T = ma, \tag{1}$$

where y stands for the desired CV inferred by the sensor, m is the vector of available input variables, and $a \in \mathbb{R}^{n_p}$ represents the vector of sensor parameters.

2.1. Industrial FCC unit

We study a depropanizer column that is a part of an FCC unit of the refinery Slovnaft, a.s. in Bratislava, Slovakia. The column separates a seven-component feed to a C3-fraction-rich distillate and to a C4/C5-fraction-rich bottom product. Plant description is given in Mojto et al. (2020). The candidate input vector for inferring the bottom impurity is:

$$m = \left(F, R, Q_{\mathrm{B}}, p_{\mathrm{D}}, p_{\mathrm{B}}, T_{\mathrm{D}}, T_{\mathrm{B}}, T_{10}, T_{37}, \frac{R}{F}, \frac{Q_{\mathrm{B}}}{F} \right), \tag{2}$$

with feed flowrate F, reflux flowrate R, reboiler heat duty Q_{B}, pressure at the top of the column p_{D}, pressure at the bottom of the column p_{B}, and temperatures of distillate T_{D}, at the 10^{th} tray T_{10}, at the 37^{th} tray T_{37} and at the bottom T_{B}. This set ($n_p = 11$) involves all variables measured directly at the column and their commonly used fractions.

Any use of a thermodynamic model to monitor top/bottom stream compositions is prohibitive here, even under some ideality assumptions. This occurs as there are too many degrees of freedom for a seven-component mixture that cannot be inferred from online data. Current inferential sensor (denoted as ref) in use in the refinery is designed according to King (2011) and uses p_{B}, T_{37}, and Q_{B}/F as inputs.

3. Soft-sensor Design by Optimal Subset Selection

This section introduces the optimal SS methods for soft-sensor design. An effective design procedure usually requires splitting the available dataset with n measurement points $\left(M := (m_1^T, m_2^T, \ldots, m_n^T)^T, Y := (y_1, y_2, \ldots, y_n)^T\right)$ into the following subsets: dataset for sensor design that contains training data $(M(\mathcal{T}), Y(\mathcal{T}))$ and dataset used for the performance evaluation of designed sensors that contains testing data $(M(\mathcal{S}), Y(\mathcal{S}))$. Here $\mathcal{T}$ and $\mathcal{S}$ denote the corresponding row-selection operators.

3.1. Optimal Subset Selection with Model-overfitting Criteria

Subset selection denotes a class of methods that explicitly seek for the simplest possible sensor structures such that some model-overfitting criterion $J(a,z)$ is minimized (Miyashiro and Takano, 2015). Here the variable z denotes a vector with binary entries $z \in \{0,1\}^{n_p}$ signifying selection of j^{th} input into sensor structure. Correspondingly, the sum of the vector entries $\sum_{j=1}^{n_p} z_j = 1^T z$ denotes the sensor complexity.

Optimal subset selection solves the following bi-level program (Bertsimas et al., 2016):

$$\min_{a,z\in\{0,1\}^{n_p}} J(a,z) \tag{3a}$$

$$\text{s.t. } a \in \arg\min_{\tilde{a}} \|Y(\mathcal{T}) - M(\mathcal{T})\tilde{a}\|_2^2 \quad \text{s.t. } -\bar{a}z_j \le \tilde{a}_j \le \bar{a}z_j, \forall j \in \{1,\dots,n_p\}, \tag{3b}$$

where $\bar{a}$ represents an upper bound on $\|a\|_\infty$ to be tuned and the optimization criterion $J(\cdot)$ might take the form (RSS $:= \|Y(\mathcal{T}) - M(\mathcal{T})a\|_2^2$):

$$J_{R^2_{\text{adj}}} = \frac{\text{RSS}}{n - 1^T z - 1} \text{ or } J_{\text{AIC}_\text{C}} = 2^T z + \log^n \frac{\text{RSS}}{n} \text{ or } J_{\text{BIC}} = \log^{1^T z}(n) + \log^n \frac{\text{RSS}}{n}. \tag{4}$$

The bi-level program (3) can be effectively solved by standard MIQP solvers using big-M reformulation as shown in Takano and Miyashiro (2020).

3.2. Optimal Subset Selection with Cross-Validation Criterion

The principle of this method is to mimic a standard cross-validation procedure within the training dataset. Let us divide the training data into K smaller subsets $\mathcal{N}_k$, such that:

$$\mathcal{T} = \bigcup_{\forall k\in\{1,\dots,K\}} \mathcal{N}_k, \quad \mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset, \ \forall k \neq k', \quad K \ge 2. \tag{5}$$

The data is distributed into training ($\mathcal{T}_k$) and validation ($\mathcal{V}_k$) sets as follows:

$$\mathcal{V}_k := \mathcal{N}_k, \quad \mathcal{T}_k := \mathcal{T} \setminus \mathcal{N}_k, \quad \text{card}(\mathcal{T}_k) \ge n_p, \ \forall k \in \{1,\dots,K\}. \tag{6}$$

where $\mathcal{V}_k$ sets contain unique data, while the different $\mathcal{T}_k$ sets involve recurring measurements. The optimal SS with cross-validation solves (Takano and Miyashiro, 2020):

$$\min_{a^{(k)},\forall k\in\{1,\dots,K\},\ z\in\{0,1\}^{n_p}} \sum_{k=1}^{K} \|Y(\mathcal{V}_k) - M(\mathcal{V}_k)a^{(k)}\|_2^2 \tag{7a}$$

$$\text{s.t. } \forall k \in \{1,\dots,K\}: \ a^{(k)} \in \arg\min_{\tilde{a}} \|Y(\mathcal{T}_k) - M(\mathcal{T}_k)\tilde{a}\|_2^2 \tag{7b}$$

$$\text{s.t. } -\bar{a}z_j \le \tilde{a}_j \le \bar{a}z_j, \ \forall j \in \{1,\dots,n_p\}. \tag{7c}$$

The problem (7) can be solved for several values of K — considering constraints on parameter identifiability, i.e., the cardinality condition in Eq. (6) — and for different randomly generated distributions of data into $\mathcal{T}_k$ and $\mathcal{V}_k$ sets. The structure of the resulting sensor is then given by the most frequent inputs occurring in the calculated sensors. Once the optimal sensor structure is calculated, a least-squares fitting of such model is used with the entire training dataset to determine the parameters of designed soft-sensor. Similarly to problem (3), the problem (7) can be effectively resolved by standard MIQP solvers.

4. Results

Industrial data available from the refinery represents more than two years of production. We possess 177 lab measurements of the bottom product composition.

We use MATLAB, Yalmip (Löfberg, 2004), and Gurobi (Gurobi Optimization LLC, 2020) to solve various instances of the problems (3) and (7). When determining the best sensor

Table 1: Comparison of the number of inputs (n_p^*) and sensor accuracy (RMSE) for the soft sensors designed using time series data.

	R^2_{adj}	AIC_C	BIC	Cross-validation	ref
n_p^*	9	4	4	4	3
RMSE	0.110	0.106	0.106	0.106	0.128

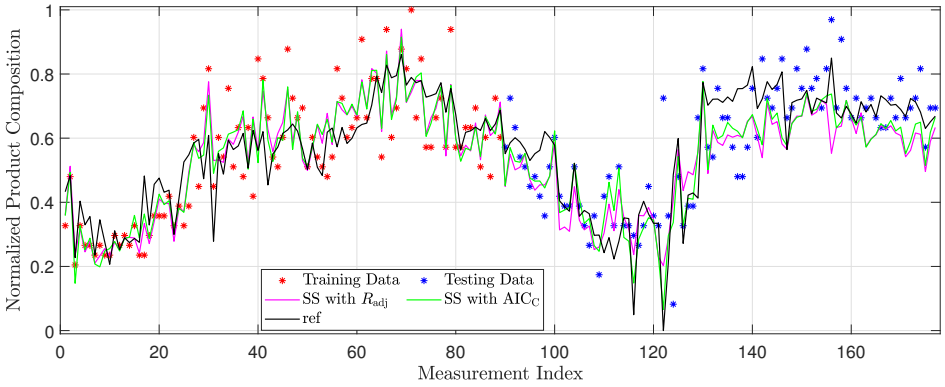

Figure 1: Comparison of the soft sensors designed using time series data.

according to SS with cross-validation, we take a median of $1^T z$ from the results of different runs (different $K \leq 6$ and data distribution) to obtain the $n_p^* \leq n_p$, i.e., the number of inputs of the final sensor. We then select the n_p^* most frequent inputs from the results of different runs to finalize the sensor structure.

4.1. Design of an Inferential Sensor using Time Series Data

We chronologically assign first 50 % of the data to the training set and the last 50 % of the data to the testing set. The accuracy of the designed sensors is assessed by root mean squared error (RMSE) evaluated on the testing dataset.

Table 1 shows the obtained results. The SS with R^2_{adj} suggests to include almost all available inputs (except T_D and Q_B) and the accuracy of this sensor is slightly decreased compared to other designed soft sensors. Therefore we conclude that this criterion is not appropriate option for structure selection of an inferential sensor. The performance of SS using AIC_C, BIC and cross-validation is the same. These methods suggest to include four common variables (T_B, T_{10}, T_{37} and Q_B/F) into the structure of the inferential sensor. Suitable candidates for a good quality soft-sensor thus seem to be temperatures in the column (T_{10} and T_{37}) and variables measured near to the inferred variable (T_B and Q_B/F).

The inferential sensors designed via SS with AIC_C, BIC, and cross-validation show better performance compared to the reference inferential sensor. We can thus conclude that an improvement of the current inferential sensor is possible with only slight modifications, i.e., at least one extra variable in the inferential sensor is required. The accuracy improvement achieved by the inferential sensor of SS with AIC_C, BIC and cross-validation is more than 15 % compared to the reference inferential sensor.

Table 2: Comparison of the number of inputs (n_p^*) and sensor accuracy (RMSE) for the soft sensors designed using randomly distributed data.

	R^2_{adj}	AIC_C	BIC	Cross-validation	ref
n_p^*	8	6	5	5	3
RMSE	0.107	0.107	0.109	0.111	0.127

Figure 1 shows a comparison of the measured data with the output of the sensors. We plot the predictions of the reference sensor and of the sensors designed by SS with R^2_{adj} and SS with AIC_C (the same as the rest of SS-based sensors). The performance of the designed sensors on the training data is good as can be expected, despite we can clearly observe problems of the reference sensor in fitting the data. This already suggests its inappropriate structure. This is further documented when looking at the testing data, where the quality of the reference sensor rapidly deteriorates once leaving the training-data window. The last period of the testing data (measurements 130–177) shows a significant discrepancy between the measurements and values inferred by all the designed sensors. This might be caused by a major change in the operating conditions of the FCC unit. A possible remedy could be to design a new sensor (with different structure) or a simple bias correction, which seems to be more appropriate in this case. The bias correction strategy is actually used at the refinery to improve the reference sensor.

4.2. Design of an Inferential Sensor using Randomly Distributed Data

To further investigate the problem, we study the impact of the training/testing data distribution on the sensors performance. Therefore, we randomly assign 50 % of the available data to the training set and leave the rest of the data for testing.

Table 2 shows the results averaged over fifty different randomly generated training/testing data distributions. These results show slightly increased complexity of the inferential sensors designed by SS with R^2_{adj} and AIC_C compared to inferential sensors designed by SS with BIC and cross-validation. Nevertheless, these SS approaches suggest five common variables (Q_B, T_B, T_{10}, T_{37} and Q_B/F) into the structure of the inferential sensor.

The sensors accuracy (see Table 2) confirms better performance of the inferential sensors designed via SS compared to the current inferential sensor. However, only SS with R^2_{adj} improved its accuracy compared to the sensors in Sec. 4.1. Therefore, the overall improvement of the designed soft sensors is comparable with Sec. 4.1 (around 15 %).

In comparison to the results shown in Fig. 1, the performance of the designed inferential sensors using random distributed data (Fig. 2) is slightly improved in the section represented by the measurements 130–177. Nevertheless, the performance of the reference inferential sensor is almost the same as in Sec. 4.1. Therefore, we can conclude that the simple structure of the reference inferential sensor provides the robustness and constant accuracy at the whole time interval. On the other hand, the designed inferential sensors with enhanced structure (more input variables) are more accurate than the reference inferential sensor, but only within a short time horizon. We thus conclude that a design of one efficient global sensor at the whole time interval is too complicated or impossible and one should better design a family of switching sensors or an appropriate mechanism for update of sensor parameters (beyond the simple bias correction). These are the directions

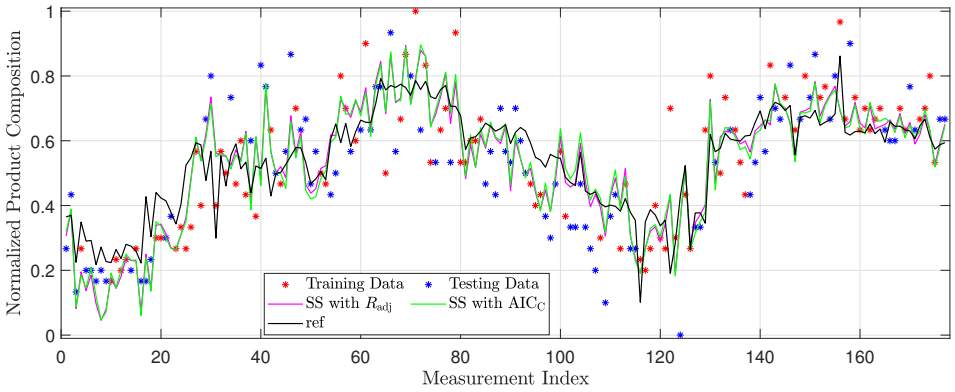

Figure 2: Comparison of the soft sensors designed using randomly distributed data.

for our further research.

5. Conclusions

We analyzed effectiveness of optimal subset selection to design a soft sensor. We used several variants of the SS method with different model-overfitting criteria and with cross-validation. According to the time series data, the use of AIC_C, BIC and cross-validation results in better performing sensors than if R^2_{adj} is used. The designed soft sensors via SS could improve the current soft sensor by around 15 %. Further investigations revealed that any further improvements would be possible using a set of switching sensors.

References

D. Bertsimas, A. King, R. Mazumder, 2016. Best subset selection via a modern optimization lens. Ann. Statist. 44 (2), 813–852.

F. Curreri, S. Graziani, M. G. Xibilia, 2020. Input selection methods for data-driven soft sensors design: Application to an industrial process. Information Sciences 537, 1–17.

Gurobi Optimization LLC, 2020. Gurobi optimizer reference manual. URL http://www.gurobi.com.

S. Khatibisepehr, B. Huang, S. Khare, 2013. Design of inferential sensors in the process industry: A review of Bayesian methods. Journal of Process Control 23, 1575–1596.

M. King, 2011. Process Control: A Practical Approach. John Wiley & Sons Ltd.

A. Kordon, G. Smits, A. N. Kalos, E. Jordaan, 2003. Robust soft sensor development using genetic programming. Data Handling in Science and Technology 23, 69–108.

J. Löfberg, 2004. Yalmip: A toolbox for modeling and optimization in MATLAB. In: Proceedings of the CACSD Conference. Taipei, Taiwan.

T. Mejdell, S. Skogestad, 1991. Composition estimator in a pilot-plant distillation column using multiple temperatures. Industrial & Engineering Chemistry Research 30 (12), 2555–2564.

R. Miyashiro, Y. Takano, 2015. Mixed integer second-order cone programming formulations for variable selection in linear regression. European Journal of Operational Research 247, 721–731.

M. Mojto, K. Ľubušký, M. Fikar, R. Paulen, 2020. Advanced process control of an industrial depropanizer column using data-based inferential sensors. Vol. 48 of Computer Aided Chemical Engineering. Elsevier, pp. 1213–1218.

S. J. Qin, T. A. Badgwell, 2003. A survey of industrial model predictive control technology. Control Engineering Practice 11, 733–764.

Y. Takano, R. Miyashiro, 2020. Best subset selection via cross-validation criterion. TOP 28, 475–488.