Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Data based multivariate pseudo correlation analysis in steel industry for optimized variable selection

Andrea Schrems*. Kurt Pichler. Konrad Krimpelstätter***. Luigi del Re*.**

**Institute for Design and Control of Mechatronical Systems, Johannes Kepler University, Linz, Austria (e-mail: andrea.schrems@jku.at, luigi.delre@jku.at).*
***Linz Center of Mechatronics GmbH, Linz, Austria (e-mail: kurt.pichler@lcm.at).*
****Siemens VAI Metals Technologies GmbH & Co, Linz, Austria (e-mail: konrad.krimpelstaetter@siemens.com)*

Abstract: Data driven variable selection, without including physical knowledge, is an important prerequisite for many applications in the field of data based modeling. This paper deals with a novel approach to optimize the dimension of the input space by a combination of common variable selection methods with multivariate correlation analysis. The results are input structures with revised pseudo correlations between input channels and a physically better interpretable structure. The presented method is successfully applied to measured data from steel industry. Some exemplary results are shown in this paper.

1. INTRODUCTION

Complex industrial systems are often equipped with extended data recording facilities, so that huge quantities of measured data are available describing the production process. These data can be used for various purposes. In some cases they are only protocol data which are stored into archives. But in a growing field of applications these measured data are used for information extraction to gain more knowledge about the underlying process, in order to improve process models, the realization of system monitoring or control optimization (Shi and Skelton (2000)).

A data based approach proves especially for industrial systems because of their complexity. Often not even process experts are aware of all relationships of their systems and able to understand every behaviour. The effort for physical modeling is mostly not reasonable and so data based modeling is chosen. In the presented application area one big challenge is the fact that such processes typically are rich of data but poor of information, since the plants are often kept in a few operating points and an arbitrary excitation is not possible due to cost concerns. The consequence of this fact is often an ill-conditioned problem and the need for methods to improve the solvability conditions.

However, independent of the final goal of an analysis, e. g. system monitoring or causal structure analysis, one important point in data based input channel selection is the dimension reduction of the solution space for the approximate solution. By using statistical methods for input selection the dimension is only bounded above by the dimension of the measured channel set and below by zero. Even for the same modelled channel the dimension is not unique, the result depends always on the used variable selection method.

To find relationships within measured data Gertler (2005) used principal component analysis (PCA) for modeling with a significantly reduced dimensionality by an orthogonal linear transformation of the data to a new coordinate system, in fact a creation of new variables with no physical meaning. George and Foster (2000) presented an empirical Bayes variable selection by using a priori distribution for the coefficients of a normal linear regression model. A nonlinear regression approach is shown by Yu *et al* (2007). They optimized nonparametric noise estimation and used it for analysing financial data. A survey of several two dimensional and multivariate variable selection methods is given by Guyon and Elisseeff (2003). Bühlmann (2007) presented an application on molecular biology. Here a further variation of a linear variable selection method for high dimensional data is discussed. All of these methods deal with the proper dimension estimation of the selected variable space, mostly in combination with maximizing the achievable model quality, to get data based models with optimized input structure.

In this paper we present a further reduction of dimensionality by elimination of pseudo correlations between selected input variables, which often leads to significantly improved input structures of the used models. We introduce a combination of different variable selection methods with multivariate correlation analysis that in fact is a step toward causal modeling like Pichler and Schrems (2007) showed.

The paper is organized as follows. Section 2 gives a short problem description and considers two dimensional and multivariate correlation methods. Section 3 introduces the

method for optimizing the dimensionality of the input space. Two exemplary results of a multivariate pseudo correlation analysis of a temper rolling mill are shown in section 4 and finally our conclusions are given in section 5.

2. PROBLEM DESCRIPTION

The overall goal of this paper is to generate data based models out of complex systems measurement data, in this case steel plants. Denote with $y = (y^{(1)}, y^{(2)}, ..., y^{(N)})$ the target variable of this measuring system with *N* observations, thus the variable to be modelled. All other measurement channels will be denoted with $\tilde{x}_1, ..., \tilde{x}_n$ whereas $\tilde{x}_i = (\tilde{x}_i^{(1)}, \tilde{x}_i^{(2)}, ..., \tilde{x}_i^{(N)})$ for $i = 1, ..., n$. These channels are the possible input variables (e. g. the possible independent variables in regression models) of the model. So the modelled channel $\hat{y}$ for the target variable y is of the form

$$\hat{y} = f\left(x_1, ..., x_m\right), \tag{1}$$

whereas $f : \mathbb{R}^{N \times m} \to \mathbb{R}^N$ is unknown just like $x_i \in \{\tilde{x}_1, ..., \tilde{x}_n\}$ for $i = 1, ..., m$ and $m \le n$.

Assume one gets perfect model parameter estimates from modeling the target variable using all input variables. Then it would be rather simple to obtain the appropriate input variables x_i by inversion of the model f. As measuring systems of steel plants tend to be data rich, but relatively information poor, the modeling process is a strongly ill-conditioned problem which is hard to handle. So the first step of the algorithm should not be the modeling, but the optimization of the solution space dimension, basically the number of input variables. This will be achieved in this paper by applying both common variable selection and multivariate correlation analysis. From the physical point of view this means to identify those variables, which are actually associated with the target variable y. The variables with no or just an indirect dependence to the target variable should not be considered in the model.

In two dimensional analysis one can compute the usual Pearson correlation coefficient to identify relationships between y and x_i, $i \in \{1, ..., n\}$. This method is not optimal for multivariate analysis as it does not consider possible influences of other variables x_j, $j \neq i$, on the correlation of y and x_i. Therefore it is not sufficient to select input variables with a high Pearson correlation to the target variable. In the literature it is recommended to use multidimensional correlation measures such as partial correlation (Garson (online), Mori and Kurata (2007) used it in power systems with a few known input variables) and part correlation (Wendorf (online)). Using these measures we are able to detect the so called pseudo correlations within the measuring system and to eliminate the according input variables from the modeling process.

Recapitulatory we use three correlation measures to optimize the dimension of a selected set of input variables:

1. Pearson correlation coefficient
2. Partial correlation coefficient
3. Part correlation coefficient

To illustrate the ideas behind those measures we make use of Venn diagrams. W. l. o. g. we assume a double regression model with the dependent variable y and the independent variables x_1 and x_2. This can be easily extended to higher dimensions. In the Venn diagram the circles represent the standardized variances of the variables (Fig. 1 and Fig. 2).

2.1 Pearson Correlation

In Fig. 1 the area (A+B+C+D=1) represents the standardized variance of y. Area B represents the part of the variance of y that can be explained by the variance of x_1, Area D represents the part of the variance of y that can be explained by the variance of x_2 and area C represents variance y which is explained by x_1 as well as x_2, so to say the intersection of variances of x_1 and x_2 in consideration of y. Squared Pearson correlation coefficient r^2_{y,x_1} is then defined as the dependence between y and x_1 regardless of all other dependencies. Thus in this illustration it is defined as

$$r^2_{y,x_1} = \frac{B + C}{A + B + C + D} = B + C\,. \tag{2}$$

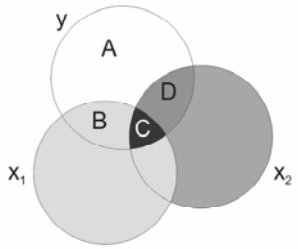

Fig. 1: Illustration of Pearson correlation

2.2 Partial Correlation

By eliminating the influence of x_2 on y and on x_1 one gets the squared partial correlation coefficient (Fig. 2)

$$r^2_{(y,x_1)|x_2} = \frac{B}{A + B}\,. \tag{3}$$

One can say it is the dependence between y and x_1 by holding all other involved variables in the set constant, both for the target channel and for the input channels.

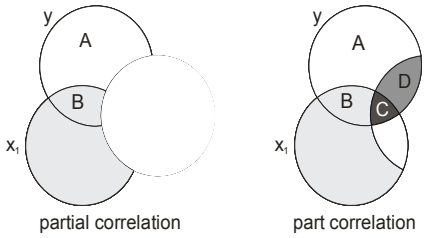

Fig. 2: Illustration of partial correlation (left hand) and part correlation (right hand)

2.3 Part Correlation

Considering just the dependence between y and x_1 by eliminating the influence of x_2 on x_1 one gets the squared part correlation coefficient (Fig. 2):

$$r^2_{y,(x_1|x_2)} = \frac{B}{A+B+C+D} = B \tag{4}$$

In a way it is the relative part of the total target channel variance that is only explained by x_1.

One problem of this correlation analysis could be that it requires several assumptions (Garson (online)):

a) linearity of relationships

b) the same level of relationship throughout the range of independent variable

c) interval or near-interval data

d) data, whose range is not truncated

Another difficulty by using these multivariate correlation coefficients is that in the literature it is recommended to use it only for small data sets. In steel plants usually a lot more than 100 variables are recorded. In order to reduce the number of input variables we execute a common variable selection step before we analyse the variables for pseudo correlations.

3. METHOD FOR DIMENSIONALITY OPTIMIZED MODEL STRUCTURES

We assume that the underlying industrial process is like a black box before data based analysis. Furthermore it is recommended to use small input variable sets for multivariate correlation analysis. Considering these two restrictions the presented method is divided into two main parts:

First: variable selection for a specified depended variable *y*

Second: multivariate correlation analysis of the variable selection result to identify pseudo correlation within the input set

In our approach pseudo correlation is defined in the following way:

3.1 Definition: Pseudo correlation criteria

The correlation between y and x_i is a pseudo correlation if both conditions are fulfilled:

$$r_{y,x_i} > 0.1 \tag{5}$$

$$mean(r_{y,(x_i|x_{1,\dots,i-1,i+1,\dots,m})}, r_{(y,x_i)|x_{1,\dots,i-1,i+1,\dots,m}}) < \varepsilon \tag{6}$$

with threshold ε sufficient small and $x_{1,\dots,m} \triangleq \{x_1,\dots,x_m\}$ representing an input variable set.

For the *first part* we are using common variable selection methods (Hennig (2004)) like

a) step forward selection linear (SFS)

b) backward selection combined with SFS linear (BSFS)

c) SFS with orthogonal projected variables in each step (LinOrth)

d) SFS with orthogonal projected variables in each step using polynomial independent variables (PolyOrth).

For the *second part* the algorithmic realization of multivariate correlation analysis proves for each pair (y, x_i) the pseudo correlation criteria (5) and (6), whereas $x_i \in \{x_1,\dots,x_m\}$ (the input channel set after variable selection).

3.2 Discussion about used variable selection methods

Each of the variable selection methods has its own advantages and disadvantages and we do not want to discuss this in more detail at this point and refer the reader to the literature (e. g. Hennig (2004)). By the experience with real industrial data it is important not only to count on one favourite method for variable selection, but to have a certain range of methods to handle the different kinds of relationships and complexity of such plants. For instance, at this point we want to remind of the information gaps inside the measurement data. But nevertheless, often the dimension of the obtained input sets is not optimal in the sense of choosing the most important input channels for a target channel without losing model quality.

In fact the used variable selection methods are looking for statistical dependencies within the measurement data, with respect to already chosen variables. One should think, that the iteratively generated input sets are optimal in the sense of the best selection dependent of the selection criteria. But we may not overlook the fact, for instance for all SFS methods, if one channel is chosen, it is not possible to eliminate it afterwards, even if it is unnecessary due to the interconnected influence of a variable that is chosen later.

3.3 Discussion about the combination with multivariate correlation analysis

With this argumentation it makes sense to check the selected input variables for pseudo correlations and try to optimize the

dimension of the input set with the positive side effect that the used multivariate correlation method is able to reveal the significant dependencies in the sense of cause and effect (Garson (online)).

The definition of the pseudo correlation criteria should be considered as a first developing step. Inequation (5) describes a necessary condition for a real relationship between two variables. The bivariate cross correlation coefficient should theoretically be greater than zero, but due to numerical reasons and measurement errors, like noise, we defined ad hoc a minimum coefficient, that turns out was a good choice for our application on steel plant data. The second condition (6) is more than an implementation of partial correlation like Mori and Kurata (2007) did. During detailed testing on our special need it turned out that combining partial correlation and part correlation gives better results than using just one method. This applies especially in the case of physically interpretable structures. The reason for that can be worked out considering the following cases in Table 1.

Table 1. Structure of partial and part correlation

	Partial Corr.		Part Corr.	
Case	high	small	high	small
1	1	0	1	0
2	1	0	0	1
3	0	1	1	0
4	0	1	0	1

Partial correlation coefficient high means that the input variable explains a big part of the target channel variance, which can not be explained by the other involved channels.
Partial correlation coefficient small means that the input channel does not add much to the unexplained variance part of the target channel (unexplained by using all other involved channels).
Part correlation coefficient high means that the input variable adds a lot to the total variance of the target channel considering all other involved input channels.
Part correlation coefficient small means that the input variable adds little to the total variance of the target channel.

Case 1 in Tab. 1 represents the situation where an input variable explains a big part of the total target channel variance **and** a big part of the variance that can not be explained by the others involved input channels. For sure such an input variable is real correlated with the target channel. Case 2 describes the combination of high partial correlation but only a small contribution to the unexplainable part of y . Here it would be important to check how large the portion of the target channel variance is that can not be explained by others. If it is small, this variable is important for the model. The situation where the input variable explains a lot of the unexplained part but less of the total variance is considered in case 3. Such a variable could be important, if the unexplained part is significant. And finally in the fourth case we consider the situation that both multivariate correlation coefficients are small. These are the typical circumstances of pseudo correlation.

From these case studies and due to the real application results we concluded that it will be the best if we start by using an average multivariate correlation coefficient with a definable threshold. Therein is some potential for further developments but for this application definition 3.1 works very well. In the following section we will show the possibility to get surprisingly well optimized variable selection results in combination with satisfying model quality.

4. APPLICATION RESULTS

The method was tested on data of a temper rolling mill as seen in Fig. 3. Temper rolling represents the final cold rolling operation after annealing in the route of the production of cold rolled strip.

Fig. 3: Temper rolling mill (Siemens VAI)

For confidentiality reasons we avoid to denote the variables with their real name or their physical units. Therefore we again denote the target variable with y and the input variables with x_i. We used linear regression models to identify and validate models.

4.1 Example One

The target variable we first analysed can be seen in Fig. 4.

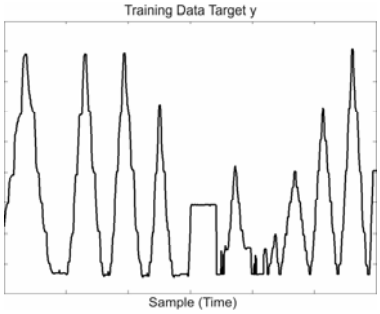

Fig. 4: Training data of the target variable

For validation we have used another data set of the temper rolling mill with the same variables. The target channel of this data set is illustrated in Fig. 5.

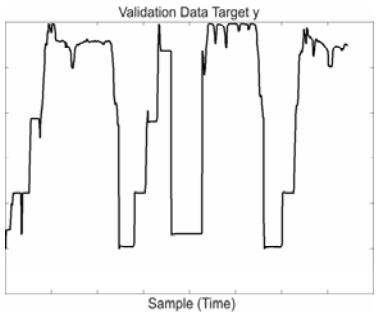

Fig. 5: Validation data of the target variable

So we have a target variable y and in the case of our measuring system 542 input variables $x_1, \ldots, x_{542}$. After applying the different methods of variable selection we obtained different input variable sets with a range between 4 and 15 input variables. The best input set we got of these was the PolyOrth set with a model validation quality of 0.9978. The corresponding model prediction compared to the validation data of the target variable is shown in Fig. 6.

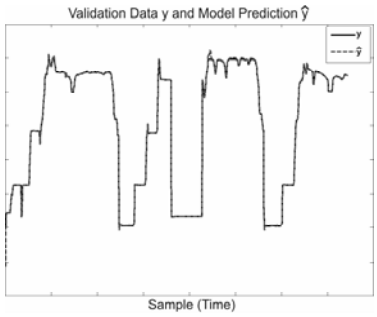

Fig. 6: Model validation for the PolyOrth input set

The worst input set (BSFS) had a validation quality of 0.5330 and is illustrated in Fig. 7.

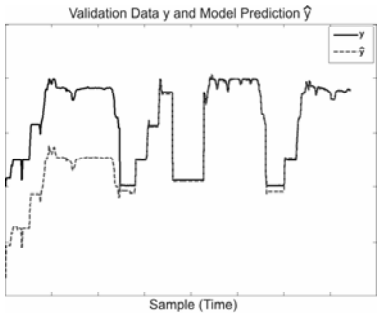

Fig. 7: Model validation for the BSFS input set

After performing multivariate correlation analysis only two input variables are left over, x_{15} and x_{16}. These two process variables, independently of the performed input variable set, were the result after pseudo correlation analysis. Building a linear regression model of y with the input variables x_{15} and x_{16} results in a validation quality of 0.9980, illustrated in Fig. 8. That is nearly the same as the best model, PolyOrth, which used 10 input variables. The selection criteria of PolyOrth were the partial F-test with a probability value of 5%. Even if we reduce the probability value to 10^{-6} % the dimension of the input space is bounded below with 4 and there is no further reduction of the dimension possible within the variable selection method.

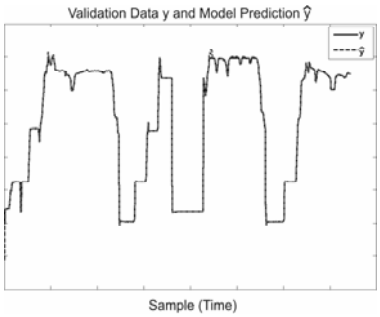

Fig. 8: Model validation for the two variable input set after multivariate correlation analysis

From process experts knowledge we know that in this case the optimized input structure of the data based model matches exactly the physical model structure. So we have shown that the dimension of the approximated input space can be optimized by multivariate correlation analysis without losing important information of the underlying process. We even achieved a gain in validation quality compared to some used variable selection methods.

4.2 Example Two

Now we chose another variable as the target (Fig. 9), the denotation is like in example one.

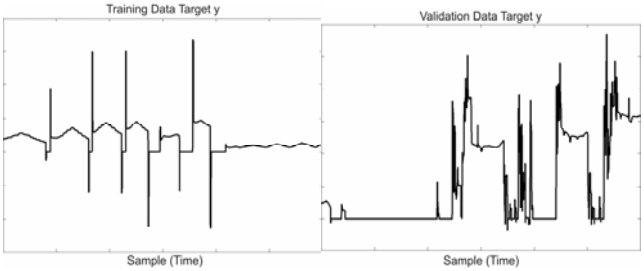

Fig. 9: Training data (left) and validation data (right) of the target variable

Performing again the presented types of variable selection we obtained different input variable sets with a range between 9 and 21 input variables. The best input set we got of these was the LinOrth set. Even this model had a poor model validation quality of 0.2976 (Fig. 10).

The worst model we obtained (SFS) had a validation quality of 0 and is shown in Fig. 11.

After applying multivariate correlation analysis to the input variable sets obtained by variable selection we got two

different sets of input variables. In two cases we got the set x_{20}, x_{524}. In the two other cases we got the input variable set x_{20}, x_{518}, x_{524}. The validation quality of these two input data sets show a negligible difference from 0.9933 to 0.9946. The validation for the two variable input set can be seen in Fig. 12.

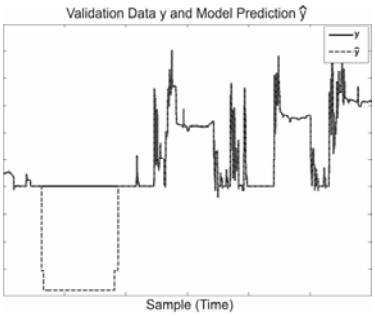

Fig. 10: Model validation for the LinOrth input set

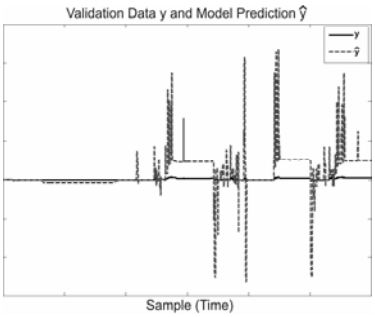

Fig. 11: Model validation for the SFS input set (the scale of the y-axis is different to Fig. 10)

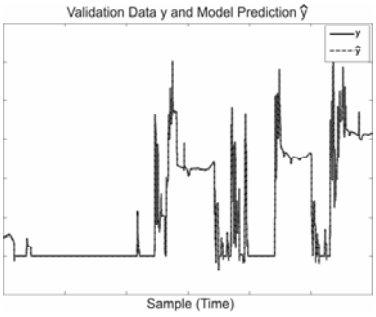

Fig. 12: Model validation for the two variable input set after multivariate correlation analysis

In this case we achieved a big gain in validation quality compared to use only variable selection by applying multivariate correlation analysis.

5. CONCLUSIONS

We presented a multivariate pseudo correlation analysis for complex systems where the reduction of the input space is an important task. Especially for black box processes with no available process expert knowledge and with missing information inside promising results could be obtained, even with very high dimensional data. Common variable selection methods represent generally a good basis for data based modeling but are often useless for process experts because the input structures frequently consist of trivial coherences between process parameters. An extensive conformity between the calculated input structures and the physical model structures strengthen the process expert that the relevant physical phenomena and influence parameters are taken into account. Mainly for systems with low or even missing process knowledge, optimized structures represent an important prerequisite to optimize process understanding and technology know-how. Consequently the next development steps should show the robustness of the method with other plant characteristics and the performance should be proved systematically for nonlinear relationships between process variables.

REFERENCES

Bühlmann, P. (2007). Variable selection for high-dimensional data. *Bulletin of the International Statistical Institute,* **56nd session.**

Garson, D. (1998). Partial correlation [Online]. Available: http://www2.chass.ncsu.edu/garson/pa765/partialr.htm (February 27, 2008).

George, E.I. and D.P. Foster (2000). Calibration and empirical Bayes variable selection. *Biometrika,* **87 (4),** pp. 731-747.

Gertler, J. (2005). Residual generation from principal component models for fault diagnosis in linear systems. *Intelligent Control, 2005. Proceedings of the 2005 IEEE International Symposium on, Mediterrean Conference on Control and Automation*, **2005**, pp. 634-639.

Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research,* **3 (2003),** pp. 1157-1182.

Hennig, C. (2004). Modellwahl und variablenselektion in der statistik [Online]. Available: http://www.math.uni-hamburg.de/home/hennig/lehre/mskript1.pdf (February 27, 2008).

Mori, H. and E. Kurata (2007). Graphical modeling for selecting input variables of short-term load forecasting. *Power Tech 2007.*

Pichler, K. and A. Schrems (2007). Data based causal modelling of manufacturing plants using transfer entropy. *IFAC MIM 2007.*

Shi, G. and R.E. Skelton (2000). Markov data-based LQG control. *Journal of Dynamic Systems, Measurement and Control*, Sept. 2000, Vol. 122, pp. 551-559

Wendorf, C.A. (1997). Multiple regression prediction with continuous variables [Online]. Available: http://www.uwsp.edu/psych/cw/statmanual/mrcontinuous.html (February 27, 2008).

Yu, Q., E. Severin and A. Bendasse (2007). Variable selection for financial modeling. *CEF 2007.*