Leakage Detection by Adaptive Process Modeling

Jaakko Talonen, Miki Sirola and Jukka Parviainen

***Abstract*— In this paper, we propose an adaptive linear approach for time series modeling and steam line leakage detection. Weighted recursive least squares (WRLS) method is used for modeling. Interpretive variables of an adaptive model should be linearly correlated to ensure a robust model. In this paper it is ensured by examining eigenvalues and eigenvectors of the principal component analysis (PCA). The method is applied to a time series from the boiling water reactor (BWR) type nuclear power plant. Model is updated and used each time step to detect leakage in steam lines. Developed leakage detection index is based on the model estimation error. Method is more convincing in small pipe flows, because there are other ways to detect bigger volume leakages.**

I. Introduction

The goal of the time series modeling presented in this paper is to detect leakage in Olkiluoto BWR type nuclear power plant (NPP) in Finland. Flow rates of the main steam lines in primary circulation system of the plant were modeled. The main feature of this approach is the use of an estimation error of the model.

In an industrial process there can be hundreds or even thousands of signals measured. How to select relevant variables for the leakage detection models? Manual selection is not possible, because of the high dimensionality of the system. Variables to the model are selected automatically when there exists a large amount of process signals [1].

Interpretive variables of an adaptive model should be linearly correlated to ensure a robust leakage detection model. Model coefficient will be zero in time, if there is no correlation between interpretive and dependent variable. When the correlation properties of the variables change with time, model has to adapt faster, and then there is a risk that small leakages are not recognized.

K-means clustering can be used to classify variables. It is unsupervised learning algorithm, which classifies a given data set through a certain number of k clusters [2]. Size of the groups are different and each object is assigned only to one group. These are the reasons, why principal component analysis (PCA) is used. It is a linear transformation to a new lower dimensional coordinate system, while retaining as much as possible of the variation. It is shown that it is also a useful method to find relevant variables for the model. One of its advantages is that the size of similar objects (number of interpretative variables) can be determined exactly. Another advantage is that same object can be selected as interpretative feature in separate models.

Jaakko Talonen, Miki Sirola and Jukka Parviainen are with the Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, 02015 TKK, Finland (phone: +358-9-451 3267; fax: +358-9-451 3277; email: talonen@cis.hut.fi, miki@cis.hut.fi, parvi@cis.hut.fi).

This work is a part of NoTeS project (Nonlinear temporal and spatial forecasting: modeling and uncertainty analysis). In this project a generic tool set for spatiotemporal forecasting and forecast uncertainty analysis is developed and it consists of five different test cases. This test case combines neuro-computing, in particular Self-Organizing Map (SOM), and time series modeling for decision support at a control room of the NPP [3].

In the second section of the paper, we describe data preprocessing, interpretative variable selection by PCA and the weighted recursive least squares (WRLS) method, used here as a modeling algorithm. In the third section, we show it's application to Olkiluoto nuclear power plant. Leakages can be detected on-line by monitoring developed *leakage index* feature.

II. Main Results

Variables in the data set should be normalized both across range and in distribution [4]. Therefore most of the multivariate methods do not work properly, if variable normalization or scaling is not done. In academic papers often used scaling method is normalization to zero mean and unit variance [1], [5], [6]. If variable has no real changes in time, only noise, normalization to unit variance increases proportional effect and creates unwanted effect to the model. Secondly, result is not very reliable, if only one data set is used for scaling, because data set can be just an exception. In this paper all variables are range scaled from zero to one.

Interpretive variables of an adaptive model should be linearly correlated to ensure a robust model. Data is projected to subspace by placing the first N principal components to matrix Θ as in

$$\Theta = (\theta_1 | \ldots | \theta_N).$$

Less significant $(N+1 \ldots m)$ directions are ignored. PCA is often used for visualizing high dimensional data in 2D or 3D. It is assumed that data is mainly concentrated in only a few directions [1], [7]. Matrix Θ and eigenvalues are used as a criteria for the variable selection. The dynamically similar variables are detected inspecting eigenvalues and eigenvectors. Similar dynamical behavior between variables has approximately same *row values* in the matrix of eigenvectors [8].

Distance between two vectors [9] $\mathbf{u}$ and $\mathbf{v}$ is defined in

$$d(\mathbf{u},\mathbf{v}) = \|\mathbf{u},\mathbf{v}\| = \sqrt{(u_1 - v_1)^2 + \ldots + (u_N - v_N)^2}.$$

For similar variables the distance is small. Weighted Euclidian distance for each variable is defined in

$$d_W(\mathbf{u},\mathbf{v}) = \sum_{k=1}^{N} \lambda_k \cdot \|\mathbf{u},\mathbf{v}\|,$$

where **u** and **v** are projected variables and N is the number of principal components in the new low-dimensional space. Weighting parameter for each component k is the eigenvalue λ_k. The idea is to stress *variable selection* in the most important dimensions.

Distance matrix is

$$\mathbf{D} = \begin{pmatrix} d_W(\mathbf{u}_1, \mathbf{u}_1) & d_W(\mathbf{u}_1, \mathbf{u}_2) & \cdots & d_W(\mathbf{u}_1, \mathbf{u}_n) \\ d_W(\mathbf{u}_2, \mathbf{u}_1) & d_W(\mathbf{u}_2, \mathbf{u}_2) & \cdots & d_W(\mathbf{u}_2, \mathbf{u}_n) \\ \vdots & \vdots & \ddots & \vdots \\ d_W(\mathbf{u}_n, \mathbf{u}_1) & d_W(\mathbf{u}_n, \mathbf{u}_2) & \cdots & d_W(\mathbf{u}_n, \mathbf{u}_n) \end{pmatrix},$$

where n is the total number of variables. It is a matrix containing the distances of set of points in new principal component space. It is a symmetric $n \times n$ matrix containing pairwise Euclidian distances. Distance matrix **D** is sorted and the variable indices of the nearest distances are captured to the matrix $\mathbf{I}_D$.

A system is static, if its statistical properties do not vary with time [10]. Complicated stationary models are used in training simulators. These are useful for training operators and understanding dynamics of the plant, but for fault detection they are not accurate enough. Industrial processes are nonstationary. For example, dependencies between variables in different process states can vary. Also external conditions such as seasonal variations impose a dynamic variation to the process. It is almost impossible to separate the effects of specific phenomena. Therefore it is hard to create static model for an industrial process. Adaptive modeling gives possibility to model and recognize abnormal events in dynamic processes. Most of the recursive models can be represented as

$$\beta(k+1) = \beta(k) + \gamma(k)(y(k+1) - \widehat{y}(k+1|k)),$$

where $\gamma(k)$ is updating factor function and $(y(k+1) - \widehat{y}(k+1|k))$ is model estimation error [11].

The most correlated variables are selected for weighted recursive least squares (WRLS) model. *Difference* vector values fits better to Gaussian distribution than original measurements. The exponential window to forget data will be wider (forgetting factor on the average is higher) and model does not adapt for leakage too fast. Therefore *difference* values of selected preprocessed variables are used as interpretative model features.

III. Simulation Results

In this work all available data sets (37) were explored, and minimum and maximum values of each variable were stored to own database of developed data management tool (DMT). Global mimimum and maximum of each process variable were used to scale process measurements into the range from zero to one. After range scaling data was converted to zero mean, because model interpretative variable selection by PCA method.

The objective of this section is to test the presented method on one stored data set, which had enough dynamics, stored variables and samples. Target variable for modeling is one of four steam lines (*steam flow 4*) in the primary circulation system of Olkiluoto boiling water reactor (BWR) in Finland. Model estimates difference value of flow rate.

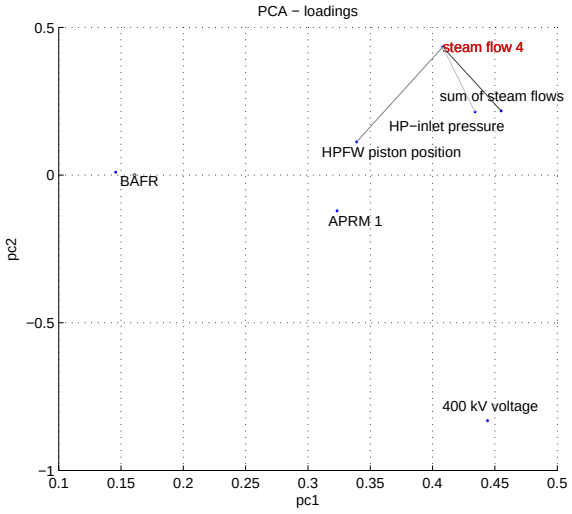

Fig. 1. A loading plot revealing relationships among the variables. Lines between the target variable and other variables illustrate the dynamical similarity. Selected interpretative variables are connected to *steam flow 4*. In this visualization only few of 70 variables are shown to ensure clarity.

After model preparation suitable interpretative variables were found by PCA, see Figure 1. Potential interpretative variables are *steam flow 1-3*, *high-pressure turbine-inlet pressure*, *high-pressure turbine feed water (HPFW) piston position*, *APRM*, *LPRM* and *feed water flow*. *APRM* and *LPRM* are neutron flux measurements in the reactor. Derived variable *sum of steam flows* was created. Different interpretative variable combinations were tested and models were evaluated in this paper simply by a mean of the squared estimation errors. Coefficient vector values $\beta(1)_i$ were inversions of the number of model interpretative features and forgetting factor $\lambda(1)$ was λ_{min}.

DMT consists of a simple algorithm to create artificial leakage for steam lines, because there exist no such data sets. Many assumptions were made for leakage, such as effects to the other process variables. Only target measurement values *steam flow 4* and values of *sum of steam flows* were manipulated. In this experiment leakage flow is about 0.04% of *steam flow 4* rate. In the end of data set total leakage volume is the same as a flow volume in *steam line 4* in a half second. Leakage starts, when 75% (t=470) of data set has been used to adapt the model coefficients, shown in Figure 2. It is logical that the correlation between process variables change in real leakage situation and information about how these dependencies exactly change is not needed in this method.

Leakage can be detected straight from Figure 3 by simply comparing difference of estimated and filtered measurements. Leakage index is simply cumulative estimation error, which is estimated leakage volume in proportion to the target variable (*steam flow 4*). It was developed for operator use and for automatic alarms. Index is reset to zero, when the estimate and filtered steam flow values intersect.

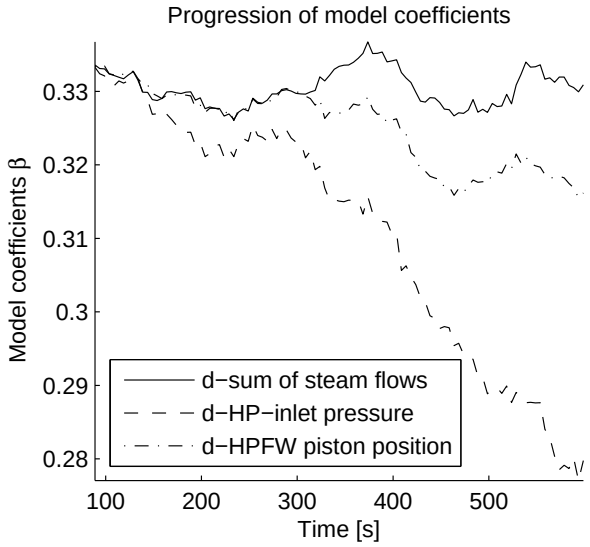

Fig. 2. Adaptive model coefficient vector values β. Range of β-values depend of model parameters. Variance of estimation error in normal condition $\sigma_v^2 = 5.0 \cdot 10^{-6}$. Typical choices of forgetting factor are in the range between 0.98 and 0.995 [11]. Here $\lambda_{max} = 1$ and $\lambda_{min} = 0.985$.

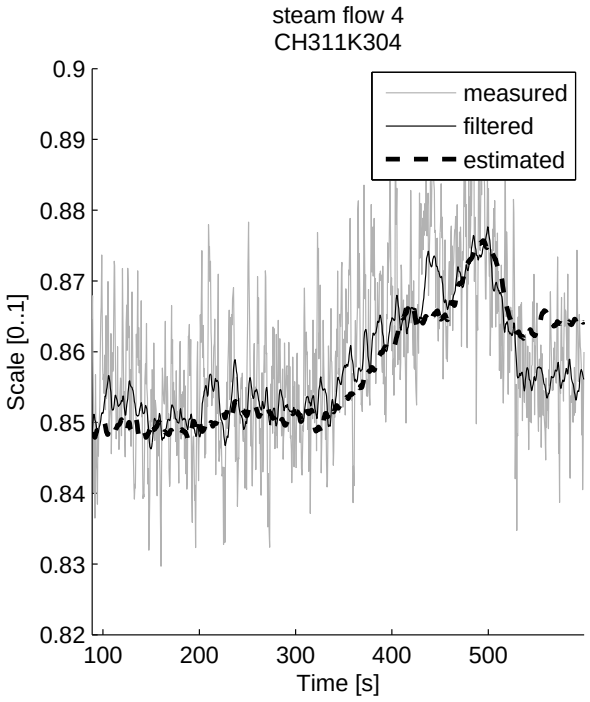

Fig. 3. Model with artificial leakage. Leakage starts at $t = 470$. *Steam flow 4* is estimated by using three interpretative features: *sum of steam flows, high-pressure turbine-inlet pressure* and *high-pressure turbine* feed water (HPFW) piston position. All data vectors were range scaled by database information of global minimum and maximum values of each variable.

The sum of squared estimation errors is multiplied by *total number of samples* and divided by *samples in subset*. These values are shown in Table I. Model coefficients are adapted in time, so leakage detection method works better in the end of data set. The proportion of cost function values is $5 : 3$ in normal process state.

TABLE I
COST FUNCTION VALUES

	$t < 470$	$t > 470$	Proportion
No leakage	0.0122	0.0073	$5 : 3$
Leakage	0.0121	0.0360	$1 : 3$
Proportion	$1 : 1$	$1 : 5$	

In Figure 4 index values are shown for leakage situation. Thin line is cumulated artificial leakage flow (volume), and wide line is estimated leakage volume. Maximum *leakage index* value without leakage was 0.07. In practice, if it is wished to avoid false alarms the limit of *leakage index* should be settled to 0.5 (Estimated leakage volume is the same as the flow in steam line in a half second). Of course many things affect to selection of the alarm limit: The criticality of the process part, dynamical properties of the process, how well WRLS model parameters are optimized and preprocessing. Leakage would not have been detected without filtering, because of the process signals noise.

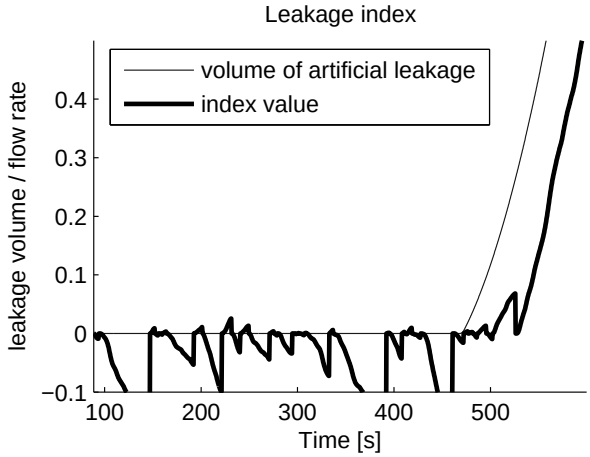

Fig. 4. *Thin line* is the proportional artificial leakage volume to steam flow 4. *Wide line* is estimated *leakage index* value. For example, with 0.2 alarm limit fault would been detected in 90 seconds $(t = 560)$

In this data set leakage can be detected fast enough. Leakage volume is the same as the flow in steam line in a half second. Problem in this application is that steam flow rates are really high in primary circulation system, about 300 kg/s. Actually, if a leakage occurs, hot high-pressure steam would fulfill the area quickly. This is detected, because there exists moisture meters in pipelines at Olkiluoto. If moisture is not detected, the steam condensation begins immediately. Water flows to a floor drain and level meters will detect the leakage.

IV. Conclusion

Principal component analysis was used for interpretative variable selection despite of its original meaning. Different adaptive models can be created easily for any process variable. In this industrial application models were created for all steam flows. Models were evaluated by cost function to find the best number of interpretative variables for each model. Leakage index is used for fault detection and it is based on cumulative estimation error of the model. It detects leakages in the steam lines in primary circulation system of boiling water reactor. The proposed method is alternative technique to detect leakages in real plant especially in small pipes, because modeling results are based on the scaled values. Leakage detection for less critical parts than primary circuit have been planned. For example channel flow lines at the power plant have about flows of 10 kg/s. So it is possible to detect leakage by process data before other sensors sense it.

References

[1] S. Laine. *Using visualization, variable selection and feature extraction to learn from industrial data*. PhD thesis, Helsinki University of Technology, 2003.

[2] S. Theodoridis. *Pattern Recognition*. Academic Press, 2003.

[3] R. Ritala, E. Alhoniemi, T. Kauranne, K. Konkarikoski, A. Lendasse, and M. Sirola. Nonlinear temporal and spatial forecasting: modeling and uncertainty analysis. *MASI Technology Programme 2005 - 2009 Yearbook*, page 10, 2007. (NoTeS) - MASIT20.

[4] D. Pyle. *Data Preparation*. Morgan Kaufmann, 1999.

[5] H. Imeläinen. Developing model structure for process automation use. Master's thesis, Helsinki University of Technology, 1997.

[6] P. Riihimäki. Development of Calibration by Multivariable Statistical Methods. Master's thesis, Helsinki University of Technology, 2004.

[7] Y.H. Chu, S.J. Qin, and C. Han. Fault Detection and Operation Mode Identification Based on Pattern Classification with Variable Selection. *Industrial & Engineering Chemistry Research*, 43(7):1701–1710, 2004.

[8] J. Talonen. Fault Detection by Adaptive Process Modeling for Nuclear Power Plant. Master's thesis, Helsinki University of Technology, 2007.

[9] H. Anton. *Elementary Linear Algebra 5e*. John Wiley & Sons Inc, 1987.

[10] J.J. Gertler. *Fault Detection and Diagnosis in Engineering Systems*. Marcel Dekker, 1998.

[11] L. Ljung. *System Identification Theory for the User*. 1999.