FAULT DETECTION BASED ON GAUSSIAN PROCESS MODELS
An Application to the Rolling Mill

Dani Juricic[1], Pavel Ettler[2] and Jus Kocijan[1]

[1]*Department of Systems and Control, Jozef Stefan Institute, Jamova 39, Ljubljana, Slovenia*
[2]*COMPUREG Plzen, s.r.o., 306 34 Plzen, Czech Republic*

Keywords: Gaussian process model, Fault detection, Statistical hypothesis test, Cold rolling mill.

Abstract: In this paper a fault detection approach based on Gaussian process model is proposed. The problem we raise is how to deal with insufficiently validated models during surveillance of nonlinear plants given the fact that tentative model-plant miss-match in such a case can cause false alarms. To avoid the risk, a novel model validity index is suggested in order to quantify the level of confidence associated to the detection results. This index is based on estimated 'distance' between the current process data from data employed in the learning set. The effectiveness of the test is demonstrated on data records obtained from operating cold rolling mill.

1 INTRODUCTION

It has been widely recognized that design of purposeful process model usually represents the bottleneck in the diagnostic system design (Venkatasubramanian et al., 2003).In this paper we investigate a relatively recent approach named Gaussian process model (GP), which is a non-parametric description of the nonlinear process (Rasmussen and Williams, 2006). We address two issues related to the application of GP in fault detection. First is the residual generation and evaluation, which is performed by calculating the error between the true output and that predicted by GP model. The zero/non-zero qualitative value of the residual is determined by means of the statistical hypothesis testing. Second, we address the problem of on-line assessment of model validity. The aim is to avoid possible false alarms that might occur due to incorrect association of the prediction error with process fault instead of insufficiently accurate process model. A model validity index is suggested, which quantifies the level of confidence associated with the detection results.

The paper is organized as follows. In section 2 the essentials of GP models are reviewed. In section 3 the detection procedure is derived. Section 4 focuses on derivation of the validity index. The ideas are implemented as part of the case study related to the monitoring of cold rolling mills. Finally, conclusions are drawn and perspectives for future work are given.

2 SYSTEMS MODELLING WITH GAUSSIAN PROCESSES

A detailed presentation of Gaussian process models can be found, e.g., in (Rasmussen and Williams, 2006) and some applications in, e.g., (Kocijan and Likar, 2008).

GP emerges as the result of central limit theorem (CLT). The nonlinear relationship between measured output $y(\mathbf{x}(k)) \in R$ and the regressor $\mathbf{x}(k) \in R^D$, where k is the sample number, can be expressed as a weighted sum of the eigen-functions $\Phi_i(\mathbf{x}(k))$

$$y(k) = y(\mathbf{x}(k)) = \sum_{i=1}^{m} w_i \Phi_i(\mathbf{x}(k))) + n(k)$$

where $n(k)$ is zero mean white i.i.d. noise. If w_i are zero mean i.i.d and $m \to \infty$ under certain conditions the CLT returns (Der and Lee, 2005)

$$y(\mathbf{x}(1)), \ldots, y(\mathbf{x}(N)) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \qquad (1)$$
$$\mathbf{\Sigma} = \| \sigma_{i,j} \|$$

$$\begin{aligned} \sigma_{ij} &= cov(y(i)y(j)) = C(\mathbf{x}(i), \mathbf{x}(j)) = \quad (2) \\ &= ve^{-\frac{1}{2}(\mathbf{x}(i)-\mathbf{x}(j))'\mathbf{W}(\mathbf{x}(i)-\mathbf{x}(j))} + v_0\delta_{ij} \end{aligned}$$

where $\mathbf{W} = \text{diag}(w_1, w_2, ..., w_D)$ are the 'hyperparameters' of the covariance functions $C : R^{D \times 2} \to R$ (Rasmussen and Williams, 2006), v_0 is the estimated noise variance, v is the estimate of the vertical scale of variation, D is the input dimension and δ_{ij} is Kronecker operator.

Consider a set of N input vectors $\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)]$ of dimension D and a vector of output data $\mathbf{y} = [y(1), y(2), \ldots, y(N)]^T$. Based on the data set $\mathcal{L} = \mathbf{X}$, and given a new input vector $\mathbf{x}^*$, we wish to find the predictive distribution of the corresponding output y^*. For a new test input $\mathbf{x}^*$, the predictive distribution of the corresponding output is $p(y^*|\mathbf{x}^*, (\mathbf{X}, \mathbf{y})) \sim \mathcal{N}(m(y^*), \sigma^2(y^*))$, i.e. is Gaussian, with mean and variance

$$m(y^*) = \mathbf{k}(\mathbf{x}^*)^T \boldsymbol{\Sigma}^{-1} \mathbf{y} \qquad (3)$$
$$\sigma^2(y^*) = k(\mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^T \boldsymbol{\Sigma}^{-1} \mathbf{k}(\mathbf{x}^*) \qquad (4)$$

where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}_1, \mathbf{x}^*), \ldots, C(\mathbf{x}_N, \mathbf{x}^*)]^T$ is the $N \times 1$ vector of covariances between the test and training cases and $k(\mathbf{x}^*) = C(\mathbf{x}^*, \mathbf{x}^*)$ is the covariance between the test input and itself.

The estimation of the hyper-parameters of the covariance function is done by maximizing the log-likelihood of the parameters. This can be computationally demanding since the inverse of the $(N \times N)$ data covariance matrix has to be calculated at every iteration. The cross-validation fit of predictions is usually evaluated by log predictive density error (Kocijan and Likar, 2008),

$$\text{LD} = \frac{1}{2}\log(2\pi) + \frac{1}{2N}\sum_{i=1}^{N}(log(\sigma_i^2) + \frac{e_i^2}{\sigma_i^2}), \qquad (5)$$

where y_i, $e_i = \hat{y}_i - y_i$ and σ_i^2 are the system's output, the prediction error and the prediction variance of the i-th element of output.

3 DETECTION PROCEDURE BASED ON GP MODELS

As measured output $y(k)$ and the predicted $\hat{y}(k)$ are both stochastic processes, detection will employ the *realization* of pairs $\{y(1), \mathbf{x}(1)\}, \ldots, \{y(i), \mathbf{x}(i)\}$ and $\{\hat{y}(1), \mathbf{x}(1)\}, \ldots, \{\hat{y}(i), \mathbf{x}(i)\}$ computed with GP model. The difference between the actual $y(k)$ and the predicted $\hat{y}(k)$ is referred to as *residual*.

3.1 Residual Generation with GPs

The prediction error can be set as follows:

$$\varepsilon(k) = y(k) - m(\mathbf{x}(k)) \qquad (6)$$

Intuitively, if the prediction error is low, there might be a good reason to infer that no fault affects the system. High prediction error, on the other hand, might mean either

- a fault is present in the system (e.g. instrument reading) or
- the process operation reached a region for which the process model is not appropriate, i.e. the vector $\mathbf{x}$ is far from the learning set $\mathcal{L}$.

Assume we have a vector of residuals $\boldsymbol{\varepsilon}(k)$ and the associated covariance matrix $\mathbf{S}(k)$ as follows

$$\boldsymbol{\varepsilon}(k) = [\varepsilon(k-M+1), \varepsilon(k-M+2), \ldots, \varepsilon(k)]^T \quad (7)$$

$$\begin{aligned} \mathbf{S}(k) &= E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\} = C(\boldsymbol{X}(k), \boldsymbol{X}(k)) \\ &\quad - C(\boldsymbol{X}(k), \mathbf{X}_L)\boldsymbol{\Sigma}^{-1}C(\boldsymbol{X}(k), \mathbf{X}_L)^T \end{aligned}$$

where $\mathbf{X}(k) = [\mathbf{x}(k-M+1), \mathbf{x}(k-M+2), \ldots, \mathbf{x}(k)]$, $\mathbf{X}_L \subset \mathcal{L}$ and the matrix

$$C(\boldsymbol{X}(k), \boldsymbol{X}(k)) = \| C(\boldsymbol{x}_i, \boldsymbol{x}_j), \boldsymbol{x}_i \in \boldsymbol{X}(k), \boldsymbol{x}_j \in \boldsymbol{X}_L) \|.$$

The problem we have now is to decide whether a bias f_y in predicted output $y(t)$ is present ($f_y \neq 0$) or not ($f_y = 0$).

3.2 Detection Rule based on a Statistical Test

If there is no fault in the system the distribution of $\boldsymbol{\varepsilon}(k)$ should read

$$\boldsymbol{\varepsilon}(k) \sim \mathcal{N}(\mathbf{0}, \mathbf{S}(k)) \qquad (8)$$

A bias error f_y in the predicted output result in offset in computed residual $\varepsilon(i)$. We have to choose between the null hypothesis

$$H_0 : f_y = 0$$

and the alternative

$$H_1 : f_y \neq 0$$

One rejects H_0 if the likelihood ratio is such that (Rohtagi, 1976)

$$\kappa = \frac{p_{f_y=0}(\boldsymbol{\varepsilon}(k))}{sup_{f_y \neq 0} p_{f_y \neq 0}(\boldsymbol{\varepsilon}(k))} < \tau \qquad (9)$$

Supremum in the denominator is achieved for

$$\mu = \frac{\mathbf{1}_N^T \mathbf{S}(k)^{-1}\boldsymbol{\varepsilon}(k)}{\mathbf{1}_N^T \mathbf{S}(k)^{-1}\mathbf{1}_N} \qquad (10)$$

where $\mathbf{1}_N^T = \underbrace{[1, \ldots, 1]}_{N-times}^T$

From the logarithm of the likelihood ratio test (9) (with (10) in mind), the following condition for rejecting the null hypothesis H_0 at the level of significance β follows

$$S_N(k) = \frac{|\mathbf{1}_N^T \mathbf{S}(k)^{-1}\boldsymbol{\varepsilon}(k)|}{\sqrt{\mathbf{1}_N^T \mathbf{S}(k)^{-1}\mathbf{1}_N}} > c_{1-\beta/2} \qquad (11)$$

Here $c_{1-\beta/2}$ is the significance level taken from the normal distribution at the degree of significance β.

4 MONITORING THE DETECTOR WITH A MODEL VALIDITY INDEX

The validity of the test (11) can be violated if GP model is forced to make predictions at the points $\mathbf{x}$ far from the learning set $\mathcal{L}$. In practice, it might often happen that process model, originally trained in certain operating region(s), is employed in a new operating region not included in the learning set.

In this paper we rely on the observation that sensitivity is inherent to the Gaussian process models. We focus on $\overline{y}$ i.e. the scaled noise-free mapping of input regressor $\mathbf{x}$, which is related to y as follows

$$
\begin{aligned}
y(\mathbf{x}(t)) &= \sqrt{v} \cdot \overline{y}(\mathbf{x}(t)) + n(t) \\
\overline{y}(t) &\sim \mathcal{N}(0, \tilde{C}(\mathbf{x}(t), \mathbf{x}(t)))
\end{aligned}
$$

where $\tilde{C} : R^{D \times 2} \to R$ is modified covariance function defined by

$$
\tilde{C}(\mathbf{x}(i), \mathbf{x}(j)) = \exp\left[-\frac{1}{2}(\mathbf{x}(i) - \mathbf{x}(j))^T \mathbf{W}(\mathbf{x}(i) - \mathbf{x}(j))\right]
$$

Furthermore, note also that $\mathbf{x}(i) = \mathbf{x}(j)$ implies $\overline{y}(\mathbf{x}(i)) = \overline{y}(\mathbf{x}(j))$. Assume $\mathcal{L} = \{\mathbf{x}_i, i = 1, ..., N\}$. Let the modified covariance matrix of the learning set $\mathcal{L}$ be

$$
\tilde{\boldsymbol{\Sigma}}(\mathcal{L}, \mathcal{L}) = \|\tilde{\sigma}_{ij} = \tilde{C}(\mathbf{x}_i, \mathbf{x}_j)\|, i, j \in \{1, ..., N\}
$$

Now, we start by adopting the notion of the distance between a new regressor $\mathbf{x}$ and the learning set $\mathcal{L}$. Intuitively, if $\mathbf{x} = \mathbf{x}_k, k \in \{1, ..., N\}$ then the distance should be zero.

Proposition 1. Assume the learning set $\mathcal{L} = \{\mathbf{x}_i, i = 1, ..., N\}$ and $\mathbf{x}$ is a regressor. The distance between x and $\mathcal{L}$ is

$$
\delta(\mathbf{x}, \mathcal{L}) = \tilde{C}(\mathbf{x}, \mathbf{x}) - \tilde{\boldsymbol{\Sigma}}(\mathbf{x}, \mathcal{L})'\tilde{\boldsymbol{\Sigma}}(\mathcal{L}, \mathcal{L})^{-1}\tilde{\boldsymbol{\Sigma}}(\mathbf{x}, \mathcal{L}). \quad (12)
$$

In eq. (12) the meaning of $\tilde{\boldsymbol{\Sigma}}(\mathbf{x}, \mathcal{L})$ is

$$
\tilde{\boldsymbol{\Sigma}}(\mathbf{x}, \mathcal{L}) = [\tilde{C}(\mathbf{x}, \mathbf{x}(1))...\tilde{C}(\mathbf{x}, \mathbf{x}(N))]'
$$

In the same manner as in case of a single regressor $\mathbf{x}$ one can associate distance to the covariance matrix of the predicted $\mathbf{y}(\mathbf{X})$ conditioned on $\mathcal{L}$. Actually, the result (12) is extended as follows

$$
\mathbf{D}(\mathbf{X}, \mathcal{L}) = \tilde{\boldsymbol{\Sigma}}(\mathbf{X}, \mathbf{X}) - \tilde{\boldsymbol{\Sigma}}(\mathbf{X}, \mathcal{L})'\tilde{\boldsymbol{\Sigma}}(\mathcal{L}, \mathcal{L})^{-1}\tilde{\boldsymbol{\Sigma}}(\mathbf{X}, \mathcal{L}). \quad (13)
$$

Again, if $X \subset \mathcal{L}$ then by borrowing the derivation from the Proposition 1 one can see that $D = 0$. If X is 'far' from $\mathcal{L}$, $D \to I$.

Definition 1. The validity index I of the GP model is proposed as being equivalent to the distance of a set of regressors from the learning set as follows

$$
I = \text{trace}(\mathbf{D}(\mathbf{X}, \mathcal{L})). \quad (14)
$$

5 EXPERIMENTAL RESULTS

To demonstrate the performance of the above FD scheme, a case of a cold rolling mill is addressed. In this process the output strip thickness belongs to the key process variables. Its control is not trivial and several approaches are being used to overcome the related technical problems (Ettler et al., 2007). One of them relies on exploiting additional redundancy based on available measured signals and mathematical models. Therefore, the on-line detection of faults in instrumentation and appropriate accommodation is key for efficient controller design. The estimated value is directly usable for the thickness control and for mill operators. The estimator output is in the form of the probability distribution thus providing clear information about reliability of the estimation (Ettler et al., 2007).

In order to illustrate the method proposed above, we will focus just on the relationship between thickness $H_2(k)$, $z(k)$ and rolling force $F(k)$, where k denotes the sampling instance. It reads as follows

$$
H_2(k) = f(z(k), F(k)) \quad (15)
$$

where f is an unknown function which can be described by Gaussian process model.

The set of 450 representative input data samples is used for training of the model (15).The model is validated with 110 input data samples different from those used for training.

The proposed statistical test (14) and the validity index (18) have been used to detect bias in the H_2 sensor. To illustrate the performance a simulation run consisting of four parts is presented.

First, in the period 0-430 samples the process operates in a fault-free mode. Moreover, in that period the operating region belongs to the region encompassed in the learning set. The response of the detector is seen in the figure. The window length is taken $M = 50$ samples. The process output is well predicted by the model so that the detector is indicating no-fault with low value of the validity index.

In the period 430-950 samples an offset $f_y = 100$ in H_2 sensor appears. One can see that the test statistic almost immediately crosses the threshold value thus indicating the presence of fault. The validity index stays around zero, indicating that the process is operating in validated region.

In the third period 950-1450 the process operates without fault.

At $k = 1450$ the operating point changes. The detector receives data not envisaged in the learning stage, Both the test statistics and validity index grows indicating that something unusual is going on in the
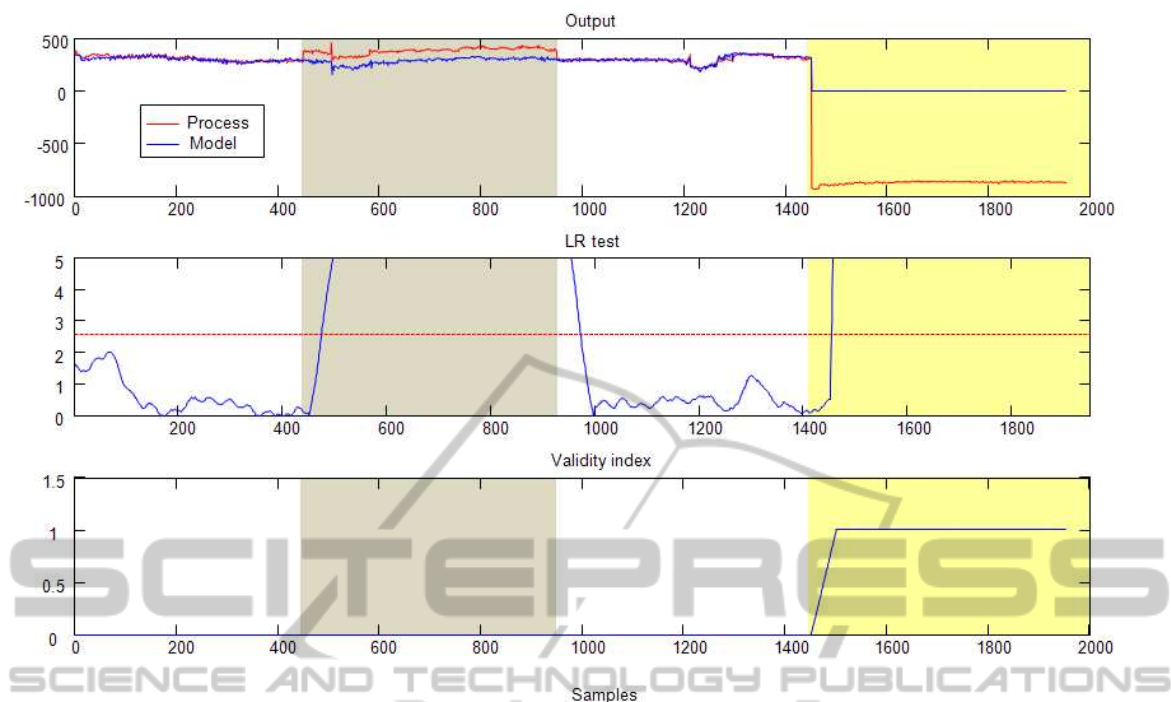
Figure 1: Results obtained over operating records.

process. Without additional information at that stage it is not possible to distinguish whether the cause is in fault or in novel operating region.

6 CONCLUSIONS

In this paper we presented a fault detection algorithm based on Gaussian process model that is suited for handling instrument faults in nonlinear systems. The main contribution of the paper regards a validity index which tells how much should the process model be trusted when deciding about faults based on data from current operating region. The idea is implemented in a Gaussian process model framework, which is suited for data-driven modelling and requires minimal a priori knowledge.Further extensions to a wider set of faults should provide more insight into the potential of the ideas presented here as well as advantages and disadvantages compared to other methods.

ACKNOWLEDGEMENTS

REFERENCES

Der R. and D. Lee , 2005. Beyond Gaussian processes: on the dist, In: *Advances in Neural Information Processing Systems*. Vol. 18, MIT Press, 275–282.

Ettler P., M. Karny and P. Nedoma , 2007. Model mixing for long-term extrapolation. *Proc. Eurosim Congress*, Ljubljana, Slovenia.

Kocijan J. and Likar, B. , 2008. Gasliquid separator modelling and simulation with Gaussian-process models, *Simulation Modelling Practice and Theory*, 16 (8), 910–922.

Rohatgi, V. K., 1976. *An Introduction to Probability Theory and Mathematical Statistics*, John Wiley and Sons, New York, NY.

Venkatasubramanian, V., Rengaswamy, R., Yin, K. and Kavuri, S. N. , 2003. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods, *Computers & Chemical Engineering*, 27(3), 293–311.

Rasmussen, C. E. Williams,C. K. I. , 2006. *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA.