CHEMICAL ENGINEERING TRANSACTIONS
VOL. 33, 2013
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at: www.aidic.it/cet
Guest Editors: Enrico Zio, Piero Baraldi

**ISBN** 978-88-95608-24-2; **ISSN** 1974-9791
DOI: 10.3303/CET1333152

# Approximate Gaussian Process Regression with Sparse Functional Learning of Inducing Points for Components Condition Monitoring

Valeria Vitelli[*,a], Enrico Zio[a,b]

[a]Chair on Systems Science and the Energetic challenge, European Foundation for New Energy - Électricité de France, École Centrale Paris, Grande Voie des Vignes, F-92295, Chatenay-Malabry, France, and Supelec (École Supérieure d'Électricité), Plateau de Moulon, 3 Rue Joliot-Curie, F-91192, Gif-sur-Yvette, France.
[b]Energy Department, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133, Milano, Italy.
valeria.vitelli@ecp.fr

We develop a novel method aimed at estimating the relationship between a drifting process parameter and some operational variables of a component under fault. The method is to be used for condition monitoring and tracking, and it is based on Gaussian Process (GP) models, which are widely used Bayesian models for nonlinear regression. These models provide great flexibility for regression and the capability of quantifying uncertainty in the form of a posterior predictive distribution. Their main limitation is the difficulty to handle large data sets. For this reason, over the past decade different approximations have been proposed to reduce the computational burden, either global or local. For global approximations of time-dependent data, a proper selection of the set of inducing points is crucial for maintaining accuracy and effectiveness, while also reducing the computational costs. In this paper, we propose a strategy to select the inducing points for nonlinear regression of time-dependent data: the algorithm adaptively computes the inducing points as sparse means over moving time-windows. The time windows are selected in order to maximize the similarity between the target variable and the inputs within each window. We finally combine the sparse functional learning of inducing point positions with an approximate GP model for nonlinear regression, with the aim of estimating the relationship between the target variable and the inputs. The effectiveness of the proposed strategy is shown on a case study with real data from a Nuclear Power Plant (NPP) component.

## 1. Introduction

In industry, methods are needed for detecting, diagnosing and controlling abnormal events in a timely and accurate manner (Venkatasubramanian, 2005). These methods can be either based on physical models, or data-driven (Zio *et al.*, 2012): the latter are particularly suited to cases in which the monitoring of the plant provides large amounts of measured data (Ma and Jiang, 2011); physical models, instead, are built and solved by simplification of the true physical relations, and in most cases this cannot timely provide the plant operators with a sufficiently precise and reliable detection and diagnostics of the plant situation (see the nice review in Zio, 2012).
Among data-driven methods for nonlinear regression, GP models are becoming popular and widely used, both for nonlinear regression and classification purposes. At the same time, they offer a flexible and powerful tool for prediction. A GP is a nonparametric technique, and as such the width of GP-based Prediction Intervals (PIs) grows in regions of the space far from the training set, where the uncertainty associated to the interpolating function is higher (Snelson, 2007). This makes the inspection of the input space a relevant aspect of the analysis, since a bad positioning of the input data can highly affect the prediction. Moreover, like all nonparametric techniques, the complexity of the model grows as more training data points are used (Snelson, 2007). Hence, the computational burden associated to GP is relevant when used for complex problems.

Please cite this article as: Vitelli V., Zio E., 2013, Approximate gaussian process regression with sparse functional learning of inducing points for components condition monitoring, Chemical Engineering Transactions, 33, 907-912 DOI: 10.3303/CET1333152

Concerning the computational aspects, the training cost for a GP has $O(N^3)$ complexity, where $N$ is the number of training data points (Storlie *et al.*, 2009). This is due to the inversion of the $N \times N$ covariance matrix. Several techniques have been recently proposed to reduce this complexity to $O(NM^2)$, where $M$ is a user-chosen number smaller than $N$. These sparse approximations to GP can be either global or local (Snelson and Ghahramani, 2006): the former try to summarize all training data via a set of $M$ representative "inducing points" (or "pseudo-inputs", since these are the only inputs employed for prediction purposes). In the case of local approximations, instead, $M$ local experts account for the prediction in different portions of the feature space. Both the selection of inducing points and the feature space partitioning are optimized together with the hyperparameters during training.
Global sparse approximations are most suited for the case of time-dependent data (Snelson and Ghahramani, 2007). In this paper, we thus focus on global approximations and develop a method for performing a careful selection of the inducing points, based on a functional similarity approach. The $M$ inducing points are selected by dividing the horizon of interest into moving windows (initially equal), and by iteratively updating these windows with the goal of maximizing the within-window similarity between the target variable and the inputs. This method is then combined with a sparse GP approach to perform the prediction of a drifting parameter of a component under fault.
The paper is structured as follows. Section 2 gives the details of the sparse approximate GP model for nonlinear regression with the functional selection of inducing points. In Section 3 an illustrative case study is introduced, concerning the condition monitoring of a NPP component under fault. Finally, in Section 4 some conclusions and future directions of research are given.

## 2. Gaussian Processes for Regression

GPs are a flexible, simple to implement and fully probabilistic approach to nonlinear regression. Consider the following nonlinear regression model

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{1}$$

where the data set $D = \{(\boldsymbol{x}_i, y_i), i = 1, \dots, N\}$ is the training set, $f$ is the latent nonlinear function describing the link between the input variables $\boldsymbol{x} \in \boldsymbol{R^p}$ and the output variable $y \in R$, and $\varepsilon_i$ is an independent additive noise term assumed with zero mean and Gaussian.
In the GP approach to regression (for a deep mathematical description of GP models, see the nice dissertation in Rasmussen and Williams, 2006), we put ourselves in a Bayesian framework assuming a Gaussian prior

$$\boldsymbol{f}|\boldsymbol{x}_1, \dots \boldsymbol{x}_N \sim N(\boldsymbol{\mu}, \mathrm{K}), \tag{2}$$

where $\boldsymbol{f} = (f_1, \dots f_N)' = (f(\boldsymbol{x}_1), \dots f(\boldsymbol{x}_N))'$ and $K$ is the $N \times N$ covariance matrix such that $K_{ij} = cov(\boldsymbol{x}_i, \boldsymbol{x}_j)$. If we consider the test point $\boldsymbol{x}^*$ with associated output $y^*$, and call $\boldsymbol{f}^* = f(\boldsymbol{x}^*)$, then in a full Bayesian approach to inference, we would put a joint GP prior on the training and test inputs, and then use the Bayes theorem to combine with the likelihood. This would lead to the following formulation

$$P(\boldsymbol{f}, \boldsymbol{f}^*|\boldsymbol{y}) = \frac{P(\boldsymbol{f}, \boldsymbol{f}^*)P(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{f}^*)}{P(\boldsymbol{y})}, \tag{3}$$

which finally allows to obtain the following predictive distribution

$$P(\boldsymbol{f}^*|\boldsymbol{y}) = N\left(K_{*\boldsymbol{f}}\left(K_{\boldsymbol{ff}} + \sigma^2 I_n\right)^{-1}\boldsymbol{y}, K_{**} - K_{*\boldsymbol{f}}\left(K_{\boldsymbol{ff}} + \sigma^2 I_n\right)^{-1}K_{\boldsymbol{f}*}\right), \tag{4}$$

where in (4) we used the shortened notation $K_{\boldsymbol{ff}}$ to indicate the matrix such that $\{K_{\boldsymbol{ff}}\}_{ij} = cov(f_i, f_j)$, and the same notation applies also to $K_{*\boldsymbol{f}}$, $K_{**}$ and $K_{\boldsymbol{f}*}$. Note that the variance of the predictive distribution in (4) needs a $O(N^3)$ computational effort, due to the presence of the training covariance matrix $K_{\boldsymbol{ff}}$. Hence, the computation of (4) is unfeasible in most high-dimensional applicative contexts, and we should introduce some proper approximations.

### 2.1 Sparse Gaussian Process approximation

The aim of the sparse approximation to GP regression is to modify the joint training and testing posterior distribution in (3), such that the resulting predictive distribution in (4) can be efficiently computed (Quinonero Candela and Rasmussen, 2005.

We, thus, introduce a set of inducing variables $\boldsymbol{u} = (u_1, \dots u_M)$, a set of latent variables which correspond to the evaluations of the GP at a set of given locations $\widetilde{\boldsymbol{x}}_1, \dots \widetilde{\boldsymbol{x}}_M$, also called inducing points. The sparse approximation consists in assuming $\boldsymbol{f}$ and $\boldsymbol{f}^*$ conditionally independent given $\boldsymbol{u}$, which means

$$P(\boldsymbol{f}, \boldsymbol{f}^*|\boldsymbol{u}) \approx P(\boldsymbol{f}|\boldsymbol{u})P(\boldsymbol{f}^*|\boldsymbol{u}). \tag{5}$$

The reason why $\widetilde{\boldsymbol{x}}_1, \dots \widetilde{\boldsymbol{x}}_M$ are named inducing points is that they "induce" the dependency between the training and the testing set, since $\boldsymbol{f}$ and $\boldsymbol{f}^*$ can only communicate through the evaluations $\boldsymbol{u}$. The exact expressions of the two conditionals are

$$P(\boldsymbol{f}|\boldsymbol{u}) = N\left(K_{\boldsymbol{fu}}{K_{\boldsymbol{uu}}}^{-1}\boldsymbol{u}, K_{\boldsymbol{ff}} - K_{\boldsymbol{fu}}{K_{\boldsymbol{uu}}}^{-1}K_{\boldsymbol{uf}}\right), \tag{6a}$$
$$P(\boldsymbol{f}^*|\boldsymbol{u}) = N\left(K_{*\boldsymbol{u}}{K_{\boldsymbol{uu}}}^{-1}\boldsymbol{u}, K_{**} - K_{*\boldsymbol{u}}{K_{\boldsymbol{uu}}}^{-1}K_{\boldsymbol{u}*}\right). \tag{6b}$$

The two approximated posterior distributions (6a) and (6b) are two noise-free versions of the predictive (4), with $\boldsymbol{u}$ playing the role of noise-free observations. This fact gives the idea of the crucial role played by the inducing points in the sparse approximation to GP regression: the set of inducing points should be capable of capturing all the relevant information contained in the inputs and, thus, to be transmitted to the output. Moreover, since the set of inducing points has dimension $M$, the computational cost of the GP regression is reduced to $O(NM^2)$ thanks to the sparse approximation.

### 2.2 Functional Learning of Inducing Points

The crucial point of the sparse GP approximation is, thus, the selection of the set of inducing points $\widetilde{\boldsymbol{x}}_1, \dots \widetilde{\boldsymbol{x}}_M$. In the present paper, a functional similarity approach is proposed, which exploits the time dependent nature of the data. Till now, we have considered a quite general nonlinear regression problem. However, in the condition monitoring applications of interest to us, the observations included in the training set (both the input and the output variables) have the characteristic of being time-dependent. Hence, we assume to observe both the inputs and the outputs on a spaced grid of $N$ time instances $t_1, \dots t_N$, belonging to the time period $[T_a, T_b]$.

Let us first introduce some notation related to our time-dependent scenario. The observations of the input variables in the training set can be written as $\boldsymbol{x}_i = \boldsymbol{x}(t_i) = (x^1(t_i), \dots x^p(t_i))'$ for $i = 1, \dots N$; analogously, the observations of the output variable can be written as $\boldsymbol{y} = (y_1, \dots y_N)' = \left(y(t_1), \dots y(t_N)\right)'$. Finally, we will use the shortened notation $y(T)$ referring to the selection $\left(y_{i_1}, \dots y_{i_r}\right)'$ of the vector $\boldsymbol{y}$ such that $t_{i_1}, \dots t_{i_r} \in T$; the same notation can be applied also to the input variables, in the corresponding multivariate setting.

To find the optimal set of inducing points, we need to introduce a $M$ dimensional set of time windows, each of them describing the region of influence of each inducing point. Then, the proposed strategy iteratively proceeds along three basic steps: for each time window, compute the weighting function maximizing the input-output similarity within the window; update the inducing points as weighted averages of the variables, given the previously computed weighting functions; finally, modify the time windows by setting the computed inducing points to be their centroid.

This procedure is rigorously described in the following scheme:

1. Initialize the $M$ dimensional set of time windows $T_1^0, \dots T_M^0$ to a uniform partition of the interval $[T_a, T_b]$, i.e. $T_m^0 \cap T_l^0 = \emptyset \ \ \forall m, l = 1, \dots M$, $\cup_{m=1}^{M} T_m^0 = [T_a, T_b]$ and $|T_m^0| = \frac{T_b - T_a}{M} \ \ \forall m = 1, \dots M$. Set $q = 0$.
2. Set $q = q + 1$, and perform the following steps:

   *Step 1.* Given the set of time windows $T_1^{q-1}, \dots T_M^{q-1}$ at the previous iteration, find the set of weighting functions $\{w_1^q, \dots w_M^q\}$ such that, $\forall m = 1, \dots M$

   $$w_m^q = \text{argmax}_{w \in L^2, ||w||_{L^2} \leq 1} \frac{1}{p} \sum_{j=1}^{p} Cov^w\left(x^j\left(T_m^{q-1}\right), y\left(T_m^{q-1}\right)\right) \tag{7}$$

   where we shortly indicated with $L^2$ the set $L^2(T_a, T_b)$, and where the weighted covariance function in (7) has the following expression

   $$Cov^w(\boldsymbol{x}, \boldsymbol{y}) = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle_w}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_w \langle \boldsymbol{y}, \boldsymbol{y} \rangle_w} \tag{8}$$

   *Step 2.* Given the set of optimal weighting functions $\{w_1^q, \dots w_M^q\}$ at the current iteration, update the set of inducing points by computing, $\forall m = 1, \dots M$

$$\tilde{\boldsymbol{x}}_m^q = \frac{\sum_{t_j \in T_m^{q-1}} w_m^q(t_j)\boldsymbol{x}(t_j)}{\sum_{t_j \in T_m^{q-1}} w_m^q(t_j)}. \quad (9)$$

Equation (9) means that the inducing points are computed as weighted averages of the inputs within the corresponding time window, a strategy which resembles the Auto-Associative Kernel Regression (AAKR) approach to prediction (see Baraldi *et al*., 2010).

*Step 3.* Given the set of optimal weighting functions $\{w_1^q, \dots w_M^q\}$ at the current iteration, update the time windows such that $\forall m = 1, \dots M$

$$t_m^q = \frac{\sum_{t_j \in T_m^{q-1}} w_m^q(t_j) t_j}{\sum_{t_j \in T_m^{q-1}} w_m^q(t_j)}, \quad (10)$$

and $t_m^q$ is the centroid of $T_m^q$.

3. Repeat 2. until the matrix norm of $|\tilde{\boldsymbol{x}}_1^q, \dots \tilde{\boldsymbol{x}}_M^q| - |\tilde{\boldsymbol{x}}_1^{q-1}, \dots \tilde{\boldsymbol{x}}_M^{q-1}|$ is smaller than a predefined tolerance $\varepsilon$.

The convergence of the procedure to a set of optimal inducing points is ensured by the well-posedness of the optimization problem (7) and by the compactness of the involved functional spaces.

## 3. Case Study

The proposed sparse GP regression method, with the associated inducing points selection strategy, is put into action in a case study concerning a set of real data from the Reactor Coolant Pump (RCP) of a NPP. In the following subsection, we describe the data and illustrate the pre-processing steps; in the subsequent one, we present the results of our approximate GP regression approach.

### 3.1 Data Description

The dataset includes the measurements of the RCP of a NPP, with increasing leak flow in the first seal (a variable included among the considered parameters). The dataset contains the values of seventeen different variables, hourly recorded along a period of 406 d, giving about 9200 time instances. The variables whose measurements concern sensors inside the RCP are nine, and they are hereafter called internal variables; the other eight are called external variables. The description of all the internal and external variables and their physical meanings are given in Table 1. In the following, the nine internal variables will be denoted with $\boldsymbol{x}^1, \dots \boldsymbol{x}^9$ and the eight external ones with $\boldsymbol{z}^1, \dots \boldsymbol{z}^8$. The target of interest for the prediction purposes is the variable $\boldsymbol{y} = \boldsymbol{x}^9$. For $t = t_{5700}$, we observe the fault, manifested by the start of a drift in the leak flow variable. We note that the data need pre-processing, because there are many outliers (bad sensor recordings) and missing observations (absence of sensors recording).

*Table 1: Physical meaning of each internal (left columns) and external (right columns) variable.*

| Variable name | Physical meaning | Variable name | Physical meaning |
|---|---|---|---|
| $\boldsymbol{x^1}$ | T cold leg loop 1 [WR] | $\boldsymbol{z^1}$ | T by-pass hot leg loop 3 |
| $\boldsymbol{x^2}$ | T water seal #1 051PO | $\boldsymbol{z^2}$ | T seal injection line |
| $\boldsymbol{x^3}$ | T stator winding motor 051PO | $\boldsymbol{z^3}$ | P primary amount file B [GL] |
| $\boldsymbol{x^4}$ | T motor lower bearing 051PO | $\boldsymbol{z^4}$ | Debit general file A |
| $\boldsymbol{x^5}$ | T lower thrust bearing 051PO | $\boldsymbol{z^5}$ | Debit general file B |
| $\boldsymbol{x^6}$ | T motor upper bearing 051PO | $\boldsymbol{z^6}$ | T aval exchanges file A |
| $\boldsymbol{x^7}$ | T motor upper thrust bearing 051PO | $\boldsymbol{z^7}$ | T aval exchanges file B |
| $\boldsymbol{x^8}$ | Flow seal injection supply RCP051PO | $\boldsymbol{z^8}$ | Debit refrigeration GMPP 051PO |
| $\boldsymbol{x^9}$ | Seal leak flow #1 RCP051PO | | |

For details on the pre-processing steps of the analysis (outlier elimination, missing data reconstruction and feature selection), we refer to Liu *et al.* (2013). In particular, the missing data reconstruction has been carried out via a local polynomial regression technique, while in the feature selection step we have

investigated the correlation structure of the dataset (among different internal and external variables), to select those variables to be included in the model. The analysis, described in Liu et al. (2013), leads to the selection, as inputs to the model, of the 6 internal variables $\boldsymbol{x}^2, \boldsymbol{x}^3, \boldsymbol{x}^4, \boldsymbol{x}^5, \boldsymbol{x}^6, \boldsymbol{x}^7$ and the 4 external ones $\boldsymbol{z}^2, \boldsymbol{z}^6, \boldsymbol{z}^7, \boldsymbol{z}^8$.

### 3.2 Results

The sparse GP regression framework has been applied considering $p = 10$ input variables, among which 6 are internal parameters of the RCP and 4 are external variables, and by estimating 20 inducing points. The tolerance for the inducing point estimation procedure has been set to $\varepsilon = 10^{-4}$. The training set is composed by observations corresponding to the first 1,000 time instances, while the testing set includes the subsequent 500 observations. The resulting weighting functions, estimated on the training set, are shown in Figure 1 (solid line): they have been plotted one after the other according to the sequence of time windows they are referred to; dashed vertical lines correspond to the boundaries of the final estimated time windows. As expected, local minima in the weighting function can be observed at the boundary between time windows. When a minimum at the boundary is not observed, as it is the case of $t = t_{800}$, this fact can be taken as an indication that the number of inducing points should be reduced. Note that the procedure automatically keeps the weighting functions nearly constant in time windows where the within-window similarity between the inputs and the target is uniform (as it is the case of the 5th, 8th, 12th and 13th time windows).

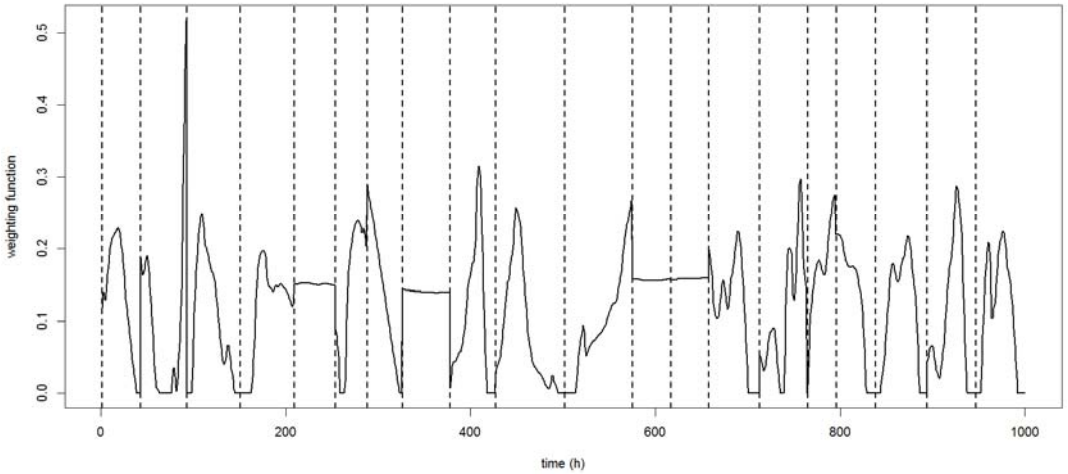


*Figure 1: weighting functions estimated by the inducing points selection procedure described in Section 2.2. Dashed vertical lines indicate the estimated time windows*

Given the optimally selected set of inducing points, we applied the sparse approximate GP method to predict the target variable $\mathbf{y}$ given the internal and external variables selected as inputs. The predictive distribution in Equation (6b) has been computed for the testing set, and the predictive distribution mean together with the 95 % prediction intervals (PIs) for the target variables have been obtained. The predictive distribution mean and the corresponding PIs are shown in Figure 2 (solid and dashed lines, respectively), together with the target variable $\boldsymbol{y}$ (dotted lines). The testing set coverage is 86 %, which is satisfactory both from the methodological and engineering perspectives.

## 4. Conclusions

GP models are flexible and widely used nonlinear regression models, capable of efficiently estimating nonlinear input-output relationships. However, their computational efficiency can become dramatically low in the context of high dimensional datasets, thus requiring sparse approximations. To achieve a global approximation, the most suited in the case of time-dependent data, a crucial point is the estimation of the set of inducing points. In the current work, we propose a novel and efficient strategy to estimate the inducing points, a set of pseudo-inputs that represents the most influencing inputs for the prediction. The strategy has solid and coherent mathematical foundations, and proves to be effective to the prediction purposes of interest.

Further developments of research will take into account different failure scenarios, describing the behaviour of the drifting process parameter of a same type of component utilized in different plants with different operational conditions.

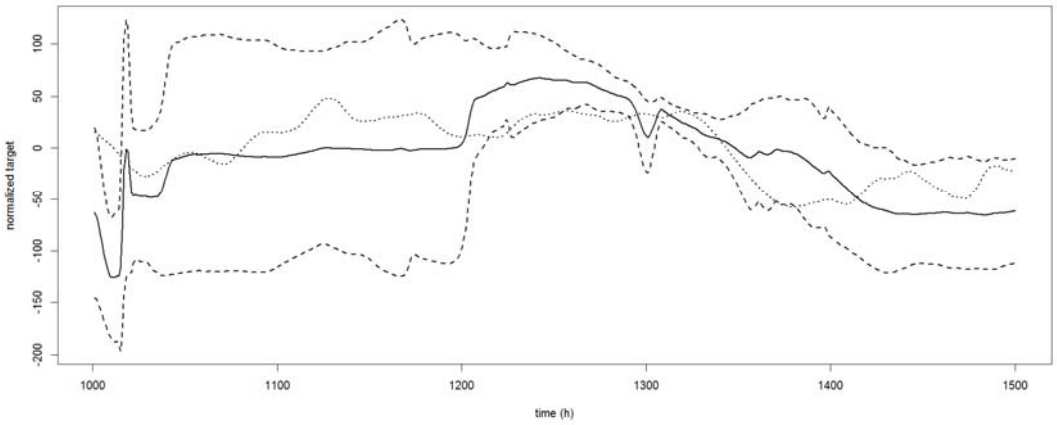


*Figure 2: mean of the predictive distribution (see Equation (6b)) of the target $\mathbf{y}$ on the testing set (solid line), estimated 95 % PIs for the target $\mathbf{y}$ on the testing set (dashed lines), and observed values of the target $\mathbf{y}$ in the testing set (dotted line)*

**Acknowledgments**

The authors are thankful to EDF R&D STEP Department for providing the data for the case study.

**References**

Baraldi P., Canesi R., Zio E., Seraoui R., Chevalier R., 2010, Signal grouping for condition monitoring of nuclear power plant components, NPIC&HMIT, Las Vegas, Nevada, USA.

Liu J., Seraoui R., Vitelli V., Zio E., 2013, Nuclear Power Plant Components Condition Monitoring by Probabilistic Support Vector Machine, Ann. Nucl. Energy, in press, DOI: 10.1016/j.anucene.2013.01.005.

Ma J., Jiang J., 2011, Applications of fault detection and diagnosis methods in nuclear power plants: A review, Prog. Nucl. Energy, 53, 255-266.

Quinonero Candela J., Rasmussen C. E., 2005, A unifying view of sparse approximate Gaussian process regression, Journal of Machine Learning Research, 6, 1939-1359.

Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning, MIT press, Cambridge, MA, ISBN 026218253X.

Snelson E. L., 2007, Flexible and efficient Gaussian process models for machine learning, PhD dissertation, University College London, London, UK.

Snelson E. L., Ghahramani Z., 2006, Sparse Gaussian processes using pseudo-inputs, Advances in Neural Information Processing Systems, vol. 18, 1257-1264, Eds. Weiss Y., Scholkopf B., Platt J., MIT press, Cambridge, MA, USA.

Snelson E. L., Ghahramani Z., 2007, Local and global sparse Gaussian process approximations, Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics, Eds. Meila M., Shen X., Omnipress.

Storlie C. B., Swiler L. P., Helton J. C., Sallaberry C. J., 2009, Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models, Reliability Engineering and System Safety, 94, 1735-1763.

Venkatasubramanian V., 2005, Prognostic and Diagnostic Monitoring of Complex Systems for Product Lifecycle Management: Challenges and opportunities, Comput. Chem. Eng., 29, 1253-1263.

Zio E., 2012, Diagnostics and Prognostics of Engineering Systems: Methods and Techniques, Chapter 17. Engineering Science Reference. USA.

Zio E., Broggi M., Golea L.R., Pedroni N., 2012, Failure and Reliability Predictions by Infinite Impulse Response Locally Recurrent Neural Networks, Chemical Engineering Transactions, 26, 117-122.