

Signal reconstruction by a GA-optimized ensemble of PCA models

Piero Baraldi, Enrico Zio, Giulio Gola, Davide Roverso, M. Hoffmann

▶ **To cite this version:**

Piero Baraldi, Enrico Zio, Giulio Gola, Davide Roverso, M. Hoffmann. Signal reconstruction by a GA-optimized ensemble of PCA models. Nuclear Engineering and Design, 2011, 241 (1), pp.301-309. 10.1016/j.nucengdes.2010.10.012 . hal-00609542

HAL Id: hal-00609542

https://centralesupelec.hal.science/hal-00609542

Submitted on 27 Jul 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Signal reconstruction by a GA-optimized ensemble of PCA models

P. Baraldi[1], E. Zio[1,*], G. Gola[1], D. Roverso[2], M. Hoffmann[2]

[1]*Department of Nuclear Engineering, Polytechnic of Milan, Via Ponzio 34/3, 20133, Milano, Italy*
[*]*enrico.zio@polimi.it*

[2]*Institutt for energiteknikk, OECD Halden Reactor Project, 1751, Halden, Norway*

Abstract

On-line sensor monitoring allows detecting anomalies in sensor operation and reconstructing the correct signals of failed sensors by exploiting the information coming from other measured signals. In field applications, the number of signals to be monitored is often too large to be handled effectively by a single reconstruction model. A more viable approach is that of decomposing the problem by constructing a number of reconstruction models, each one handling an individual group of signals. To apply this approach, two problems must be solved: 1) the optimal definition of the groups of signals and 2) the appropriate combination of the outcomes of the individual models. With respect to the first problem, in this work, Multi-Objective Genetic Algorithms (MOGAs) are devised for finding the optimal groups of signals used for building reconstruction models based on Principal Component Analysis (PCA). With respect to the second problem, a weighted scheme is adopted to combine appropriately the signal predictions of the individual models. The proposed approach is applied to a real case study concerning the reconstruction of 84 signals collected from a Swedish nuclear boiling water reactor.

1. Introduction

During plant operation, sensors may experience anomalies and convey inaccurate or misleading information on the actual plant state to the automated controls and the operators. Hence, it is important to develop accurate and robust systems capable of detecting such anomalies and correctly reconstructing the signals of the failed sensors.

On-line sensor monitoring allows detecting the degradation in the instrument performance; upon such detection, recalibration can be initiated. This way of proceeding based on the monitored health condition of the instrumentation bears important economic advantages, as the current quest for increased competitiveness by all industries requires streamlining all plant operations and reducing system downtime. In general, benefits of this procedure include the reduction of unnecessary maintenance and increased confidence in the actual values of the monitored parameters, with important consequences on system operation, production and accident management [1, 2].

The purpose of this work is to provide a robust technique for reconstructing faulty signals. In many field applications, the number of measured signals is too large to be handled effectively with one single reconstruction model [2-5]. A viable approach to address the issue of signal dimensionality amounts to subdividing the set of signals into smaller overlapping groups, developing a reconstruction model for each group of signals and then combining appropriately the outcomes of the models within an ensemble approach [6-8] (Figure 1).

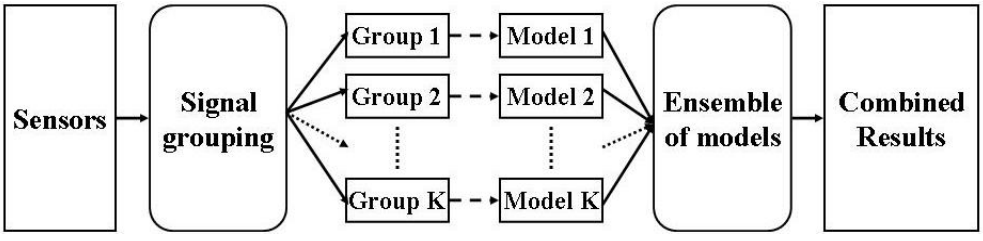


Figure 1. The multi-group ensemble approach to signal reconstruction

A major motivation behind the use of an ensemble of models is that reliance on different models increases the robustness of reconstruction. By assigning appropriate performance indicators to the individual models and basing the reconstruction on those models in the ensemble indicated as best performing for a robust reconstruction, failed or de-calibrated signals can be effectively reconstructed.

Two issues are central to the ensemble approach for signal reconstruction: (1) the partitioning of the signals into groups and (2) the combination of the outcomes of the individual models developed on the basis of the groups.

In this work, the first issue is tackled by resorting to Multi-Objective Genetic Algorithms (MOGAs) [9-14]. The groups thereby found are then used as bases for developing a corresponding number of reconstruction models based on Principal Component Analysis (PCA) [15-19]. Regarding the second issue, the outcomes of the models are combined by a weighted-combination scheme to provide the ensemble-reconstructed signal values.

The paper is organized as follows. Section 2 illustrates the MOGA-based method for signal grouping. The ensemble approach is then presented in Section 3 and applied in Section 4 to a real case study concerning the reconstruction of a data set of 84 signals measured at a Swedish nuclear Boiling Water Reactor located in Oskarshamn. Some conclusions on the advantages and limitations of the proposed methods are drawn in the last Section.

2. Multi-Objective Genetic Algorithms for Signal Grouping

The problem of signal grouping is here framed as a Multi-Objective Genetic Algorithm (MOGA) optimization search [10-14, 20, 21]. Formally, given $n >> 1$ sensors' signals f_i, $i = 1,2,\ldots,n$, the aim of the MOGA search is to group them in K groups with some required characteristics and each one constituted by $m_k < n$ signals, $k = 1,2,\ldots,K$. The inclusion or not of a signal in a group can be encoded in terms of a binary variable which takes value 1 or 0, respectively.

The MOGA probabilistic search is thus performed on a population of K chromosomes, each one constituted by n bits representing all the signals: in the generic k-th chromosome, coding signal group k, the i-th bit, s_i^k, encodes the presence (1) or absence (0) of the i-th signal in the k-th group, $i = 1,2,...,n$, $k = 1,2,...,K$ (Figure 2).

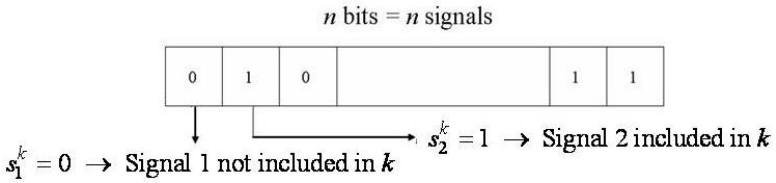


Figure 2. The structure of the generic k-th chromosome

In general, in order to effectively group the signals, the objective functions of the genetic search should capture both the individual properties of the groups (i.e., the mutual information content of the group signals [4, 5, 21]) and the global properties related to the ensemble of groups (i.e., the diversity between the groups, the redundancy of the signals and the inclusion of the majority of the signals in the groups [2, 3, 5]).

The MOGA optimization devised in this work consists of a two-objective genetic search which at convergence leads to a final population constituted by groups which maximize the group signals correlation and the diversity between the groups in the population.

The first objective function is intuitively motivated by the fact that the signals in the groups are used to build models for their reconstruction and by the conjecture that strongly positively or negatively correlated signals are capable of regressing one another. In fact, the information content of strongly negatively correlated signals is also very high and comparable to the one derived from strongly positively correlated signals. The measure herein used to quantify these characteristics is the Pearson's correlation coefficient [22, 23]. Considering N measurements of two signals $f_p(t)$ and $f_q(t)$, $t = 1,2,...,N$, the Pearson's coefficient is defined as:

$$corr_{p,q} = \frac{1}{N-1}\sum_{t=1}^{N}\left(\frac{f_p(t)-\overline{f}_p}{S_{f_p}}\right)\left(\frac{f_q(t)-\overline{f}_q}{S_{f_q}}\right) \quad (1)$$

where $\overline{f}_p$, S_{f_p}, $\overline{f}_q$ and S_{f_q} are the mean values and standard deviations of the signals f_p and f_q, respectively.

By definition, the value of $corr_{p,q}$ ranges from -1 to 1 and assumes the value of 0 for statistically independent quantities. Signals that have the same trend (both increasing or both decreasing with respect to the mean) will have positive values of $corr_{p,q}$, whereas those which vary in opposite ways will render $corr_{p,q}$ negative.

To associate a degree of correlation to the generic k-th group of m_k signals, $k = 1, 2, ..., K$, the average absolute correlation of each p-th signal, $p = 1, 2, ..., m_k$, is first computed as the mean of the absolute values of the correlation between the p-th and the remaining $m_k - 1$ signals, viz.

$$\left\langle corr_p \right\rangle = \frac{1}{m_k - 1} \sum_{\substack{q=1 \\ q \neq p}}^{m_k} \left| corr_{p,q} \right| \tag{2}$$

Finally, the group correlation r_k, computed as the geometric mean of the average correlations of the m_k signals in the group [24], is taken as first objective function for the group optimization:

$$r_k = \sqrt[m_k]{\prod_{p=1}^{m_k} \left\langle corr_p \right\rangle} \tag{3}$$

This measure allows assigning a low correlation to those groups in which at least one signal has a low average correlation with respect to the others in the group [24].

As for the second objective function, the measure used to compute the diversity between the groups is based on the concept of Hamming distance [25, 26]. According to the definition of the chromosome illustrated in Figure 2, to compute the pairwise diversity measure $div_{j,h}$ between the j-th and h-th group in the population, first the numbers of different, $diff_{j,h}$, and common signals, $com_{j,h}$, between the two groups are computed:

$$diff_{j,h} = \sum_{i=1}^{n} \left| s_i^j - s_i^h \right| \tag{4}$$

$$com_{j,h} = \sum_{i=1}^{n} \delta_{j,h}(i) \tag{5}$$

where $s_i^{j,h}$ is 1 if the i-th signal is included in the j-th or h-th group, respectively, and 0 viceversa and $\delta_{j,h}(i)$ equals 1 if the i-th bit of both the j-th and h-th group is equal to 1, i.e. if the two groups have the i-th signal in common, and 0 viceversa.

The pairwise diversity between the j-th and h-th groups is computed based only on the signals effectively included in the two groups, viz.:

$$div_{j,h} = \frac{diff_{j,h}}{m_j + m_h - com_{j,h}} \tag{6}$$

This measure is equal to 0 if two groups are constituted exactly by the same signals (i.e., $diff_{j,h} = 0$) and to 1 if the groups have no signals in common.

The group diversity, d_k, between the k-th group and the others $j = 1, 2, ..., K$ in the population is finally computed as the average pairwise diversity:

$$d_k = \frac{1}{K - 1} \sum_{\substack{j=1 \\ j \neq k}}^{K} div_{k,j} \tag{7}$$

At the end of the search, the final population is characterized by K highly correlated and diverse groups of signals; these groups are used as bases for developing a corresponding number of reconstruction PCA-based models; these models constitute the ensemble for reconstructing the signals, as illustrated in the following Section.

3. The Ensemble Approach to Signal Reconstruction

The aim of the ensemble approach is to ensure an accurate and robust reconstruction of the signals. The approach is sketched in Figure 3.

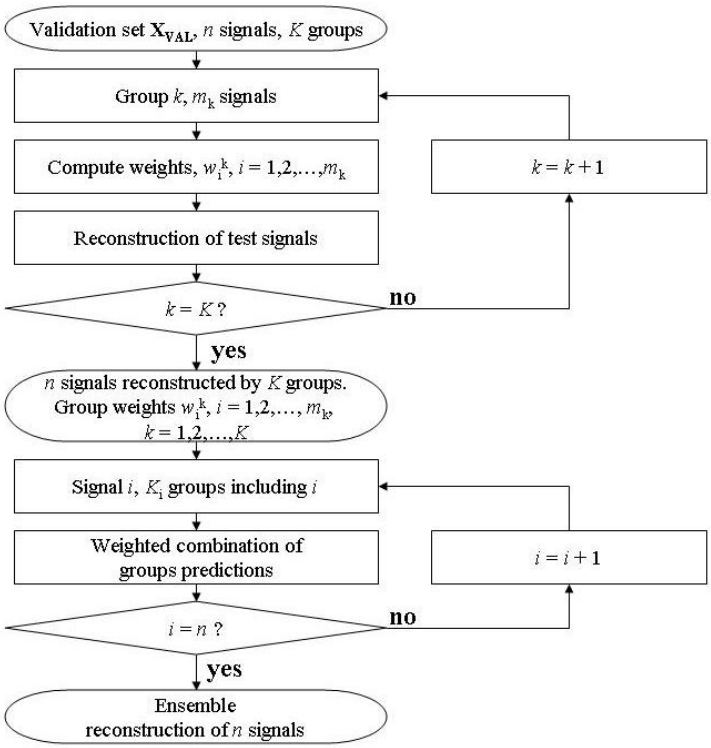


Figure 3. Sketch of the ensemble algorithm for signal reconstruction

To integrate the signal predictions provided by the K groups a weighted combination-scheme has been adopted. This way of combining the models outputs can be looked at as an extension to a regression problem of the Static Weighted Voting (SWV) technique adopted in classification problems to combine the class assignments of the single classifiers constituting the ensemble [27].

In this approach, each group k is assigned m_k weights, w_i^k, $i = 1,2,...,m_k$, the i-th weight being proportional to the accuracy of the model developed with group k in reconstructing the i-th signal. The set $\mathbf{X}_{VAL}$ of N_{VAL} patterns available to validate the ensemble performance is first partitioned into a training set $\mathbf{X}_{TRN}$ (made of N_{TRN} patterns) and a test set $\mathbf{X}_{TST}$ (made of N_{TST} patterns). The former is used to compute the sets of weights of the individual models, whereas the latter is used to verify the ensemble performance in the reconstruction task. To compute the weights, the training set must be further partitioned into a so-called train-induce set $\mathbf{X}_{TRN-IND}$ (made of $N_{TRN-IND}$ patterns) and an induce set $\mathbf{X}_{IND}$ (made of N_{IND} patterns), as illustrated in Figure 4. The weight w_i^k is proportional to the error in reconstructing the N_{IND} induce patterns of the i-th signal made by the model trained on the $N_{TRN-IND}$ patterns of the train-induce set. Thus, in the ensemble average more importance is given to the predictions of those groups whose corresponding models better reconstruct signal i.

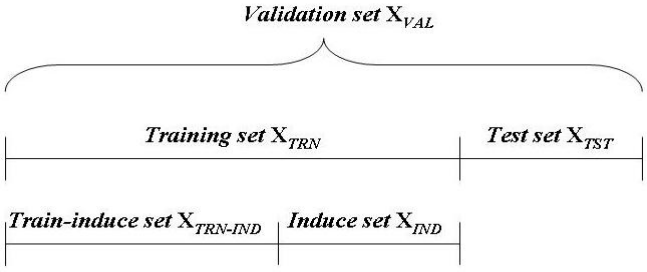


Figure 4. Subdivision of the validation set

Operatively, the weights for the generic k-th group are obtained by first computing the absolute reconstruction error for each signal $i = 1, 2, ..., m_k$ on the normalized induce data[1], viz:

$$\varepsilon_i^k = \frac{1}{N_{IND}} \sum_{t=1}^{N_{IND}} \left| f_i(t) - \hat{f}_i^k(t) \right| \tag{8}$$

where, for $t = 1, 2, ..., N_{IND}$, $f_i(t)$ is the real (normalized) value of the i-th signal and $\hat{f}_i^k(t)$ is the reconstruction obtained by the PCA model built using the signals of the k-th group and trained on the $N_{TRN-IND}$ patterns of the train-induce set.

The corresponding weight w_i^k, $i = 1, 2, ..., m_k$, is simply taken as the inverse of the signal reconstruction error:

$$w_i^k = \frac{1}{\varepsilon_i^k} \tag{9}$$

Obviously, this measure allows assigning large weights to those groups which provide the most accurate reconstruction of signal i.

Once the weights are computed, the generic PCA model built using the m_k signals of group k is trained on the N_{TRN} patterns of the training set to provide in output the predictions $\hat{f}_i^k(t)$, $t = 1, 2, ..., N_{TST}$, $i = 1, 2, ..., m_k$ of the N_{TST} patterns of the test set.

The ensemble reconstruction of the N_{TST} patterns of the generic i-th signal, $\hat{f}_i^E(t)$, $t = 1, 2, ..., N_{TST}$, is then obtained by combining the predictions $\hat{f}_i^k(t)$ of the K_i groups including signal i with the corresponding weights w_i^k, $k = 1, 2, ..., K_i$:

$$\hat{f}_i^E(t) = \frac{\sum_{k=1}^{K_i} \hat{f}_i^k(t) \cdot w_i^k}{\sum_{k=1}^{K_i} w_i^k} \tag{10}$$

Finally, to evaluate the ensemble performance, first the absolute signal reconstruction error is computed according to Eq. (8) using the ensemble signal predictions:

$$\varepsilon_i^E = \frac{1}{N_{TST}} \sum_{t=1}^{N_{TST}} \left| f_i(t) - \hat{f}_i^E(t) \right| \tag{11}$$

Then, the ensemble accuracy is retained as the average of the absolute signal reconstruction errors:

$$\eta^E = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^E \tag{12}$$

4. Application

The ensemble approach has been applied to a real case study concerning n=84 signals collected at a nuclear Boiling Water Reactor (BWR) located in Oskarshamn, Sweden (Figure 5).

The MOGA code used for the calculations has been developed by the Laboratorio di Analisi di Segnale e di Analisi di Rischio (LASAR, Laboratory of Analysis of Signals and Analysis of Risk) of the Department of Nuclear Engineering of

[1] In the application which follows, each signal of the validation data set has been previously normalized in the range [0.2, 1].

the Polytechnic of Milan (http://lasar.cesnef.polimi.it). The PCA code has been taken from http://lib.stat.cmu.edu/multi/pca and adapted to perform the signal reconstruction tasks.

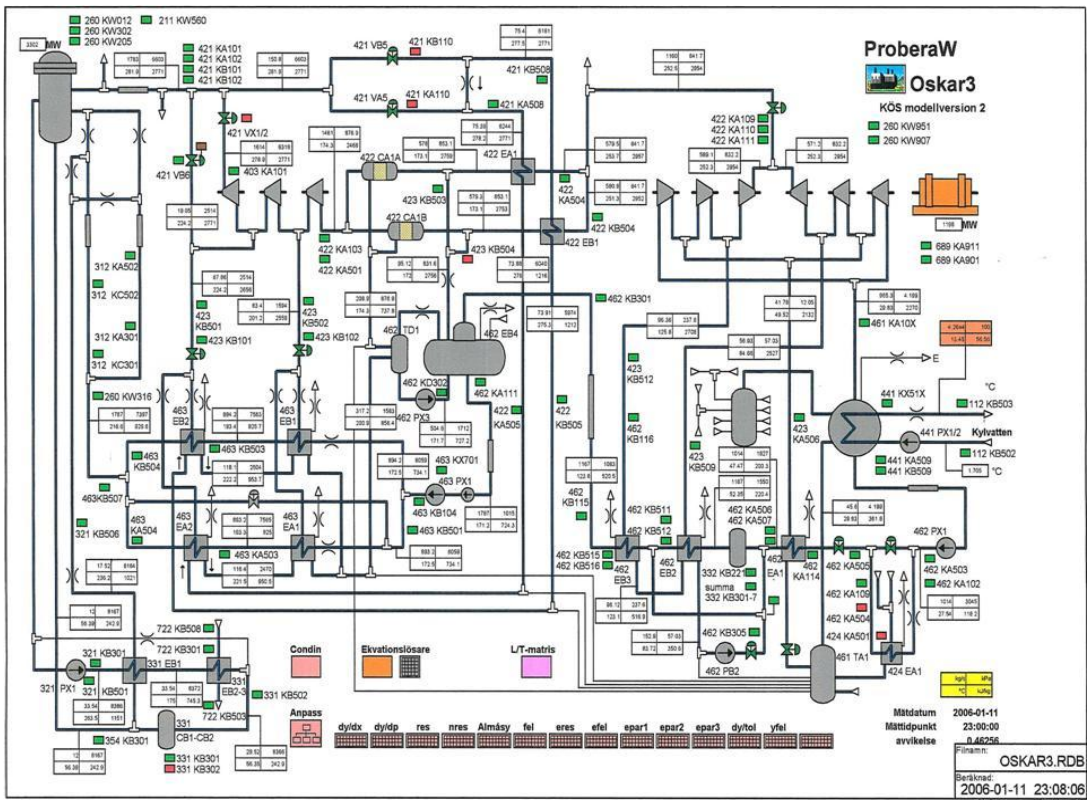


Figure 5. The process diagram of the nuclear power plant in Oskarshamn. The 84 signals are identified by an alpha-numeric code (e.g. 312 KA301, 423 KB509, etc…).

The data available for the 84 signals have been sampled every 10 minutes from May 31, 2005 to January 5, 2006 from a corresponding number of sensors, providing a total amount of $N = 30080$ time samplings. Each recording instant t provides an 84-dimensional pattern $f_1(t), f_2(t),..., f_{84}(t)$, $t = 1,\ldots, N$, identified by the values of the $n = 84$ signals at the time instant t, as illustrated in Table 1.

The total number of available patterns is divided into a MOGA set $\mathbf{X}_{MOGA}$ of $N_{MOGA} = 14966$ samples used to perform the genetic optimization and a validation set $\mathbf{X}_{VAL}$ of $N_{VAL} = 15114$ samples used to train and test the PCA models and to compute the ensemble reconstruction performances (Eqs. 11 and 12).

Time samplings $(t = 1,2,...,N)$		Signals $(i = 1,2,...,n)$			
$N_{MOGA} = 14966$	1	$f_1(1)$	$f_2(1)$	…	$f_{84}(1)$
	…				
	14966				
$N_{VAL} = 15114$	14967				
	…				
	30080	$f_1(30080)$	$f_2(30080)$	…	$f_{84}(30080)$

Table 1. Partition of the total data set for the MOGA signal grouping and ensemble signal reconstruction

4.1 MOGA signal grouping

The results of the two-objective MOGA search (Section 2) carried out to find the optimal set of groups are illustrated below. The MOGA settings here adopted are reported in Table 2, with a synthetic explanation of the terms used.

Selection	FIT-FIT: the population, rank-ordered on the basis of the Pareto dominance criterion, is scanned and each individual is parent-paired with an individual of the next fittest Pareto rank class
Replacement	FITTEST: out of the four individuals (two parents and two chromosomes) involved in the crossover procedure, the fittest two replace the parents
Mutation probability	10^{-2}
Population size	100
Number of generations	30000

Table 2. Main parameters used for the MOGA search. For further details on the FIT-FIT selection procedure and the FITTEST replacement procedure the interested reader may consult [9, 10].

At convergence of the search, $K = 100$ highly correlated and diverse groups of signals make up the final population. Figure 6 reports the main features (correlation, diversity and size) of the optimal groups of the final population. The groups contain from 2 to 70 signals and are characterized on average by a good degree of diversity. The value of diversity computed according to Eq. (7) is generally lower for small groups, which are naturally the most correlated [4]. In general, the groups in the final population are characterized by different trade-offs between the two objectives.

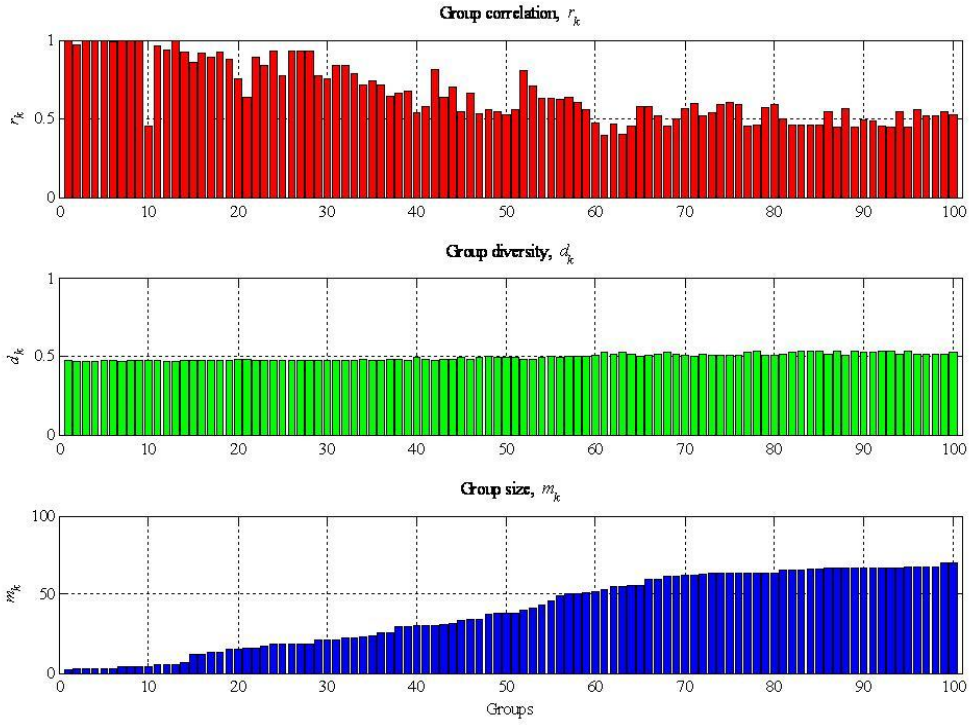


Figure 6. Main features of the $K = 100$ groups of the final population in ascending order with respect to their size. Each group is identified by the value of its correlation r_k (Eq. 3) (top), diversity d_k (Eq. 7) (middle) and size m_k (bottom).

As mentioned in Section 2, an ensemble approach to signal reconstruction is effective only if the groups in the ensemble are sufficiently diverse, they ensure a good signal redundancy and include the majority of the signals [2, 3]. Figure 7 shows that the maximization of diversity as objective function has led to fulfil both these global properties.

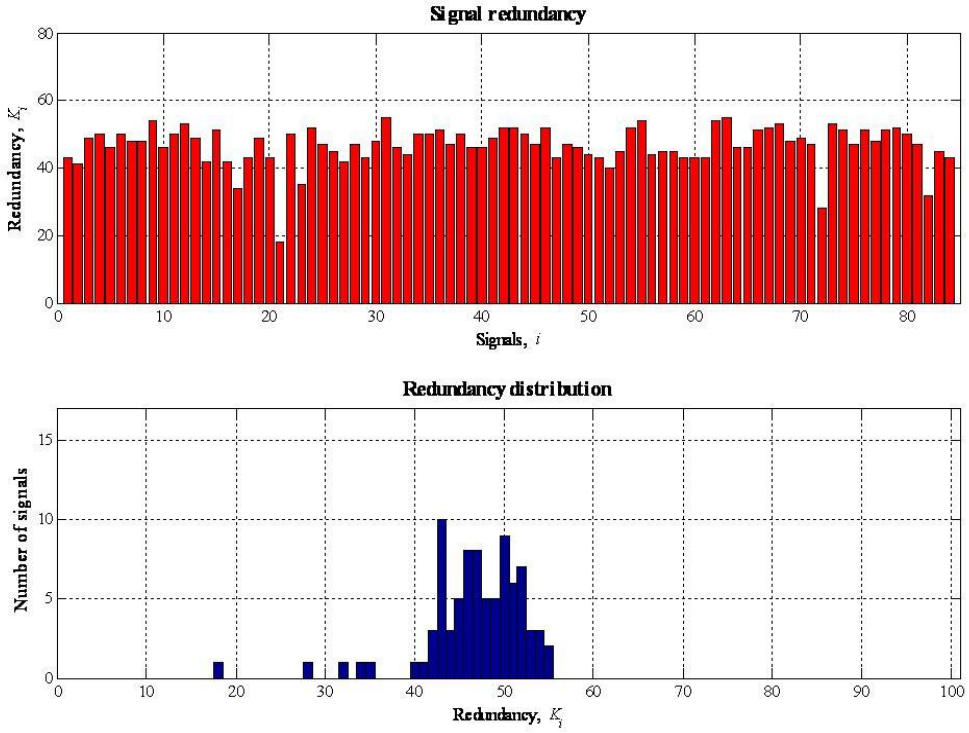


Figure 7. Redundancy of each of the $n = 84$ signals (top) and corresponding distribution (bottom)

4.2 Ensemble signal reconstruction

Within the proposed ensemble weighted approach (Section 3), the signals of the K groups are used as bases for developing a corresponding number of PCA reconstruction models, as illustrated in Figure 1.

The validation set $\mathbf{X}_{VAL}$ originally formed by $N_{VAL} = 15114$ patterns has been reduced to $N'_{VAL} = 1511$ patterns by sampling one every ten. Note that preliminary studies have shown that reducing the number of samples has negligible effects on the values of the errors, while it considerably decreases the computational cost of validating the ensemble of models.

The ensemble procedure has been 20 times cross-validated. At each run, $N_{TRN} = 1200$ patterns are randomly sampled from the $N'_{VAL} = 1511$ validation patterns to constitute the training set $\mathbf{X}_{TRN}$ needed for building the PCA models; the remaining $N_{TST} = 311$ patterns form the test set $\mathbf{X}_{TST}$ for evaluating the performance of the ensemble of models. Furthermore, a 10-fold cross-validation has been performed to compute the group weights. In this respect, at each cross-validation, $N_{TRN-IND} = 1000$ patterns randomly sampled from $\mathbf{X}_{TRN}$ constitute the train-induce set $\mathbf{X}_{TRN-IND}$, the other $N_{IND} = 200$ patterns forming the induce set $\mathbf{X}_{IND}$.

For each group, the number of principal components retained to reconstruct the test patterns has been set equal to half the size of the group, for it represents a good compromise between the accuracy and the computational cost of the model [5, 15-19].

The first row in Table 3 reports the results of the reconstruction in terms of the ensemble error η^E defined as (12). Compared with the error made by a single reconstruction model built on all 84 signals (last column of Table 3), the ensemble shows an improvement of 26.6%.

As stressed in the previous Sections, a robust ensemble of models must be able to reconstruct the signals when in presence of sensor failures, e.g. random noises, offsets or drifts. Within the proposed ensemble approach, a faulty sensor sends a faulty signal in input to the PCA models which include that signal; in this situation, the ensemble of models should still be capable of providing a good estimate of the true value of the signal by exploiting the information coming from the non-faulty signals in the groups of the ensemble.

To verify this, a disturb is introduced in the test set. More precisely, the signals of a test pattern are randomly affected either by a random noise (with probability $p^{RN} = 0.02$) or by setting their value equal to the offset value of the corresponding sensor (with probability $p^{OF} = 0.01$); with probability 0.97 they are not affected at all.

Defining γ_i as the total disturb affecting signal i:

$$\gamma_i = \sum_{t=1}^{N_{TST}} \left| \frac{f_i(t) - f_i^*(t)}{f_i(t)} \right| \tag{13}$$

where $f_i(t)$ and $f_i^*(t)$ are the real and disturbed test patterns, respectively, the average percentage amount of disturb $\langle \gamma \rangle_{\%}^{TST}$ introduced in the test set is:

$$\langle \gamma \rangle_{\%}^{TST} = 100 \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \gamma_i \right) \tag{14}$$

The magnitude of $\langle \gamma \rangle_{\%}^{TST}$ is proportional to the intensity of the random noise; obviously, $\langle \gamma \rangle_{\%}^{TST} = 0$ if no disturb has been applied to the test set.

The second row in Table 3 reports the results obtained with disturbed patterns. Once again, the ensemble of models outperforms the single model, with an improvement of 20.7 %.

However, when applying the ensemble weighted approach on disturbed signals, the computation of the group weights becomes a crucial issue. In fact, calculating the group weights as before using undisturbed training patterns may lead to large weights for groups with indeed good reconstruction performances on the undisturbed signals, but not suitable for regressing correctly the disturbed signals.

To overcome this limitation, a robust training procedure has been embedded in the ensemble reconstruction scheme. In this respect, the training patterns are disturbed in the same manner as those of the test set, with a corresponding amount of disturb $\langle \gamma \rangle_{\%}^{TRN}$ defined as in Eq. (14). By so doing, the group weights are proportional to the group reconstruction performances on disturbed signals.

In the last two rows of Table 3, the accuracy and robustness of the multi-group ensemble approach with robust training are compared to those of the single reconstruction model built with all 84 signals. Robust training is seen not to improve the robustness of the model; on the other hand, if no robust training is performed (rows 1 and 2), the reconstruction error when the test set is disturbed is significantly increased (+1680% from row 1 to row 2 against +130% from row 3 to 4).

		Reconstruction errors ($\times 10^{-1}$)	
		Ensemble, η^E	Single group
$\langle \gamma \rangle_{\%}^{TRN} = 0$	$\langle \gamma \rangle_{\%}^{TST} = 0$	0.0094 ± 0.0096	0.0128 ± 0.0151
	$\langle \gamma \rangle_{\%}^{TST} = 0.61$	0.1676 ± 0.0111	0.2114 ± 0.0138
$\langle \gamma \rangle_{\%}^{TRN} = 0.61$	$\langle \gamma \rangle_{\%}^{TST} = 0$	0.0756 ± 0.0247	0.0933 ± 0.0259
	$\langle \gamma \rangle_{\%}^{TST} = 0.61$	0.1742 ± 0.0213	0.2306 ± 0.0267

Table 3. Ensemble approach compared to a single model, with and without robust training.

To investigate further the effectiveness of robust training, larger amounts of $\langle \gamma \rangle_{\%}^{TST}$ have been considered to verify whether the advantage brought by robust training in reconstructing disturbed signals overcomes the loss of accuracy in regressing undisturbed signals. As an indicator, let $g_A(\%)$ be the absolute percentage gain obtained with robust training:

$$g_A(\%) = 100 \cdot \left(\eta^E \Big|_{\langle \gamma \rangle_{\%}^{TRN} = 0} - \eta^E \Big|_{\langle \gamma \rangle_{\%}^{TRN} = 0.61} \right) \tag{15}$$

Figure 8 shows that the effectiveness of robust training increases with $\langle\gamma\rangle_{\%}^{TST}$ and a sensible advantage is achieved only for $\langle\gamma\rangle_{\%}^{TST} \gtrsim 4$ (Figure 9).

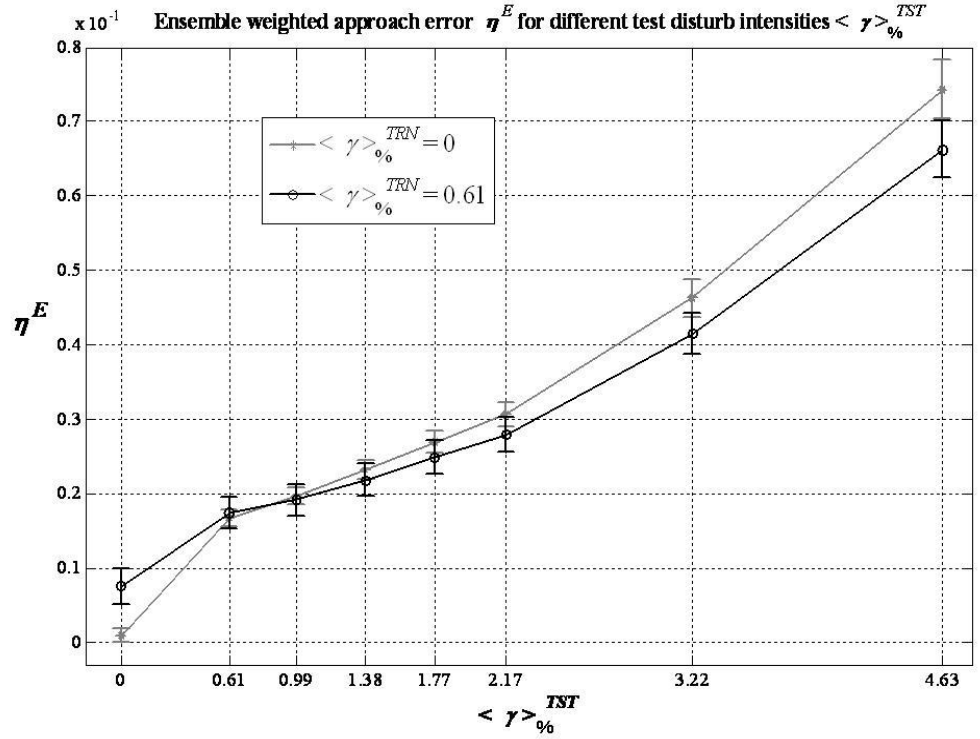


Figure 8. Ensemble reconstruction error, η^E, without and with robust training for different $\langle\gamma\rangle_{\%}^{TST}$.

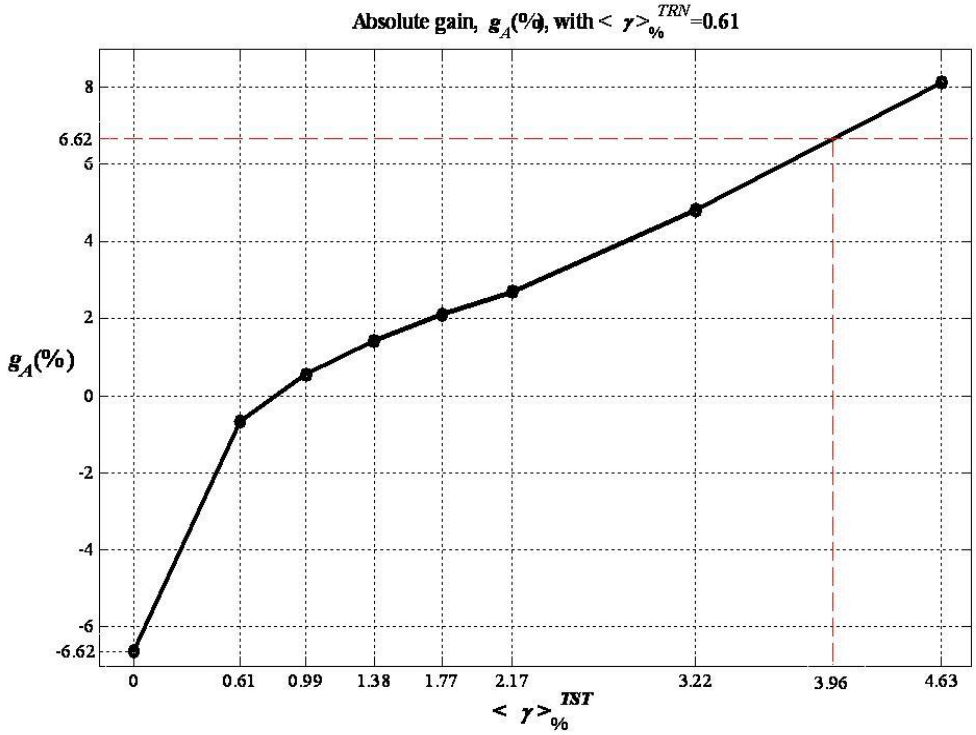


Figure 9. Absolute percentage gain $g_A(\%)$ with robust training for different $\langle\gamma\rangle_{\%}^{TST}$. The 6.62% loss of accuracy in terms of $g_A(\%)$ in regressing undisturbed signals $\left(\langle\gamma\rangle_{\%}^{TST} = 0\right)$ is effectively counterbalanced by a gain for $\langle\gamma\rangle_{\%}^{TST} > 3.96$.

The ensemble robustness has been then specifically tested on the reconstruction of a faulty signal when other sensor signals are damaged. Signal 7 has been chosen as the object of the analysis. A linear drift decreasing its value up to 25% of its real value has been introduced; at the same time, other signals have been either linearly drifted or forced in offset. The different fault scenarios are reported in Table 4.

The validation set has been linearly divided into training and test (Figure 4) only once (no cross-validation), whereas a 10-fold cross-validation has been performed on the computation of the group weights. Table 4 reports the ensemble performances on signal 7 in terms of the signal reconstruction error ε_7^E (Eq. 11), with and without robust training; Figures 10a-c show the reconstruction of signal 7 in cases A, B and D, respectively.

Faulty signals		Ensemble reconstruction error on signal 7, ε_7^E ($\times 10^{-1}$)		
		$\langle\gamma\rangle_{\%}^{TST}$	$\langle\gamma\rangle_{\%}^{TRN} = 0$	$\langle\gamma\rangle_{\%}^{TRN} = 0.61$
A	2 highly correlated linearly drifted	0.41	0.8338	0.4910 ± 0.1742
B	2 poorly correlated linearly drifted	0.39	0.4078	0.4486 ± 0.1660
C	10 randomly selected linearly drifted	1.26	0.6077	0.5187 ± 0.1813
D	1 highly correlated in offset	1.16	2.0713	0.6290 ± 0.3805
E	1 poorly correlated in offset	0.37	0.3643	0.4582 ± 0.1607

Table 4. Ensemble performance in the reconstruction of signal 7, when linearly drifted in different fault scenarios. When performing robust training, 20 differently disturbed training sets with $\langle\gamma\rangle_{\%}^{TRN} \approx 0.6$ are generated and the mean ensemble reconstruction error on signal 7 is considered.

Robust training (last column) turns out to be effective when highly correlated signals are disturbed (cases A and D in Table 4, Figures 10a and 10c, dark dots); this is true, even if $\langle\gamma\rangle_{\%}^{TST}$ is small (case A). In fact, performing robust training helps reconstructing the highly correlated signals useful to reconstruct signal 7. On the contrary, when disturbing uncorrelated signals (cases B and E in Table 4), robust training is indeed useful for correctly reconstructing the disturbed signals, but it unavoidably introduces disturbs in the reconstruction of the correlated signals effectively exploited for reconstructing signal 7, thus worsening the reconstruction accuracy (Figure 10b, dark dots). In case C, highly and poorly correlated signals have been drifted and the effectiveness of robust training is due to the high value of $\langle\gamma\rangle_{\%}^{TST}$.

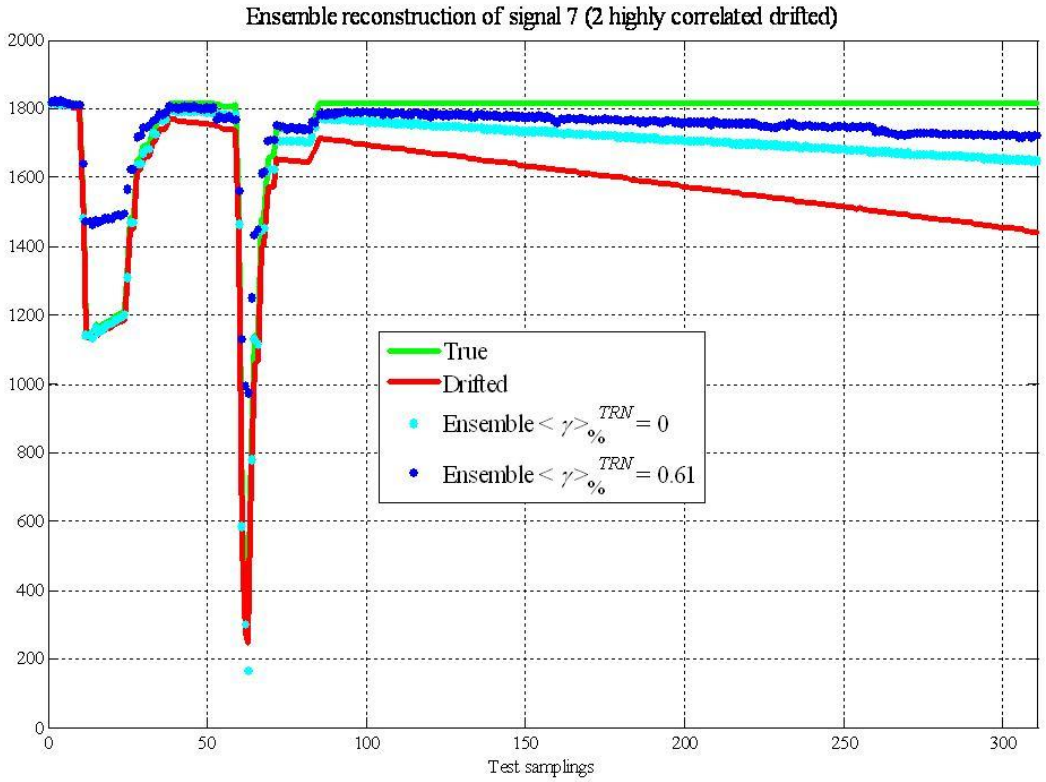


Figure 10a. Ensemble reconstruction of signal 7 (light line) when linearly drifted (dark line) in case A, with (dark dots) and without (light dots) robust training.

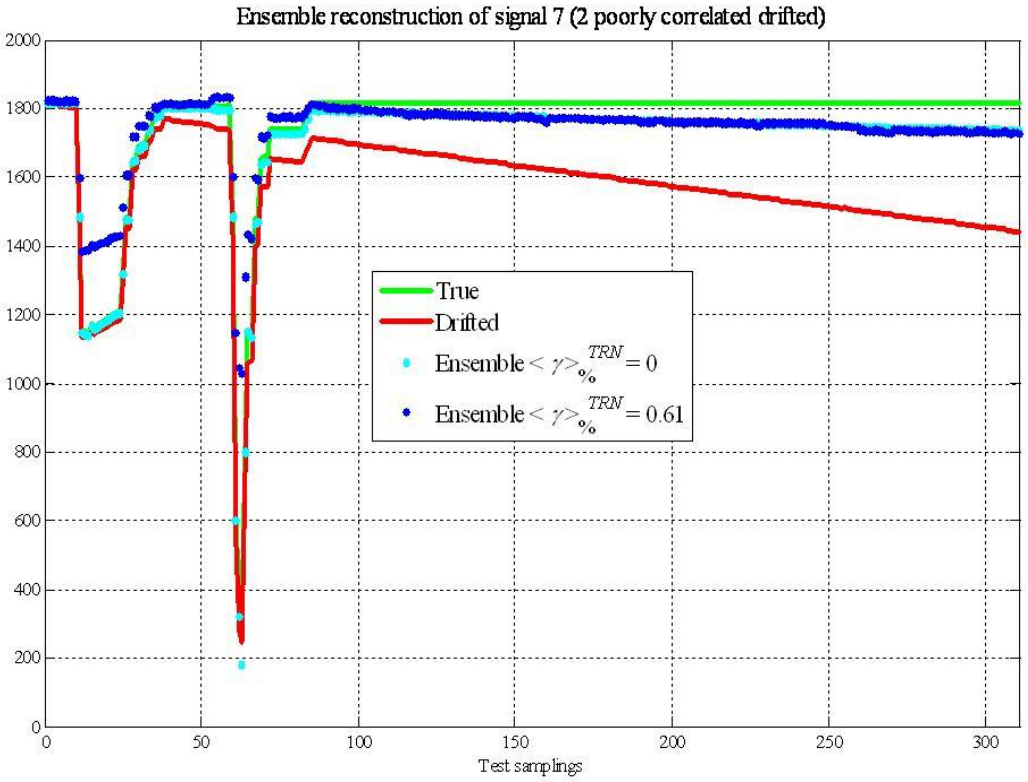


Figure 10b. Ensemble reconstruction of signal 7 (light line) when linearly drifted (dark line) in case B, with (dark dots) and without (light dots) robust training.

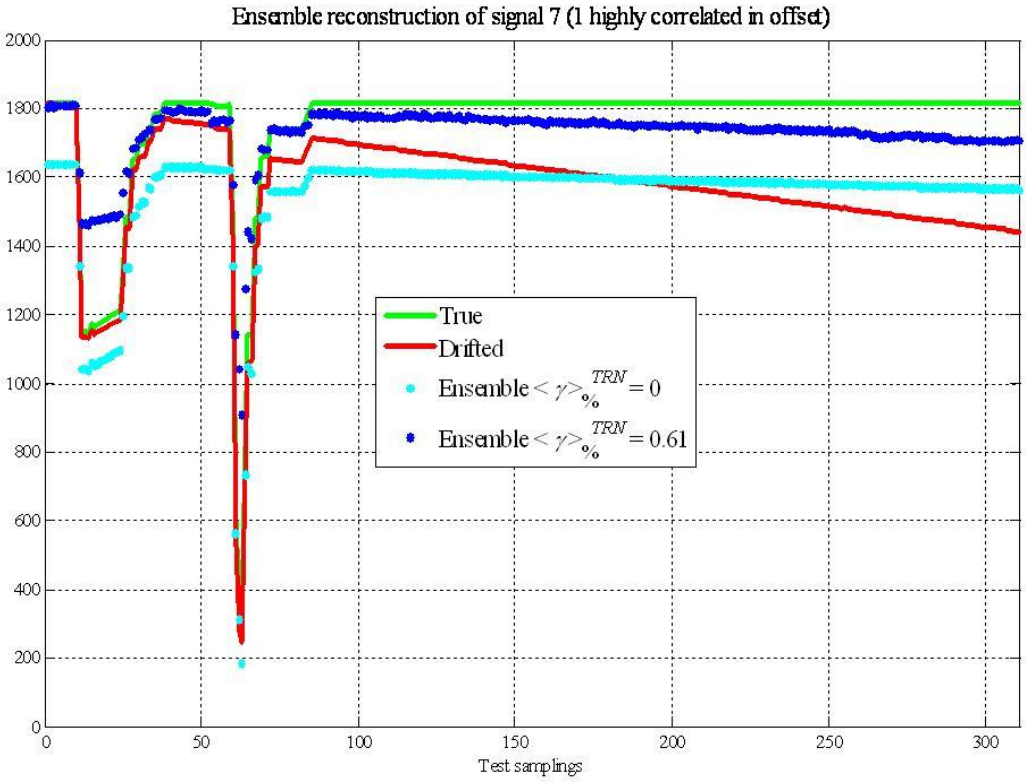


Figure 10c. Ensemble reconstruction of signal 7 (light line) when linearly drifted (dark line) in case D, with (dark dots) and without (light dots) robust training.

5. Conclusions

In this work, signal reconstruction has been carried out by resorting to a multi-group ensemble approach. The set of sensor signals, too large to be handled effectively with one single reconstruction model, is subdivided into many small, overlapping groups and a reconstruction model is developed for each group. The outcomes of the models are then combined to obtain the ensemble signal reconstruction.

Multi-Objective Genetic Algorithms have been used to generate groups of correlated signals; Principal Component Analysis has been used to build the corresponding reconstruction models; a weighted combination scheme has been used to combine the outcomes of the models.

The overall modelling scheme has been applied for the reconstruction of signals collected at a Swedish nuclear boiling water reactor. To enhance robustness, a robust training procedure has been introduced. The reconstruction performances obtained by the ensemble approach have been compared with those of a single model built using all the signals: the multi-group ensemble approach has proved to be more accurate and robust than the single model, independently of robust training.

Finally, the reconstruction capabilities with and without robust training have been tested on one faulty signal in various multi-fault scenarios. The results have shown that robust training provides a more accurate reconstruction of the signal when highly correlated signals are affected by disturbs, whereas slightly lower performances are achieved when uncorrelated signals are drifted, for the introduced disturbs actually worsen the reconstruction of the signals correlated to the one of interest, upon which reconstruction is based.

References

[1] M. Hoffmann, “On-line Monitoring for Calibration Reduction”, HWR-784, OECD Halden Reactor Project, October 2005.

[2] M. Hoffmann, “Signal Grouping Algorithm for an Improved On-line Calibration Monitoring System”, Proceedings of FLINS, Genova, Italy, Aug. 2006.

[3] D. Roverso, M. Hoffmann, E. Zio, P. Baraldi, G. Gola, “Solutions for plant-wide on-line calibration monitoring”, 2007, Proc. ESREL 2007, Stavanger, Norway, Vol. 1, pp. 827-832.

[4] E. Zio, P. Baraldi, G. Gola, D. Roverso, M. Hoffmann, “Genetic Algoritms for Grouping of Signals for System Monitoring and Diagnostics”, 2007, Proc. ESREL 2007, Stavanger, Norway, Vol. 1, pp. 833-840.

[5] P. Baraldi, E. Zio, G. Gola, D. Roverso, M. Hoffmann, “Genetic Algorithms for Signal Grouping in Sensor Validation: a Comparison of the Filter and Wrapper Approaches”, 2007, submitted for publication to Journal of Risk and Reliability (JRR).

[6] M. P. Perrone and L. N. Cooper, “When networks disagree: ensemble methods for hybrid neural networks”, 1992, National Science Fundation, USA.

[7] A. Krogh and J. Vedelsby, “Neural network ensembles, cross-validation and active learning”, in: G. Tesauro, D. S. Touretzky and T. K. Loen, 1995, Advances in newel information processing systems, Vol. 7, pp. 231-238, MIT press, Cambridge, MA, USA.

[8] A. J. C. Sharkey, “On combining artificial neural nets”, 1996, Connection Science, Vol. 8(3), pp. 299-314.

[9] J.H. Holland, “Adaptation in natural and artificial systems: an introductory analysis with application to biology”, Control and Artificial Intelligence. MIT Press, 4-th edition, 1975.

[10] D.E. Goldberg, “Genetic algorithms in search, optimization, and machine learning”, Addison-Wesley Publ. Co., 1989.

[11] L. Chambers, “Practical handbook of genetic algorithms: applications Vol. I; new frontiers Vol. II”, CRC Press, 1995.

[12] Y. Sawaragy, H. Nakayama, T. Tanino, “Theory of multiobjective optimization”, Academic Press, Orlando, Florida, 1985.

[13] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Khun, A. K. Jain, “Dimensionality reduction using genetic algorithms”, IEEE Transaction on Evolutionary Computation, vol. 4, No. 2, July 2000.

[14] H. Bozdogan, “Statistical Data Mining with Informational Complexity and Genetic Algorithm”, in: H. Bozdogan, “Statistical Data Mining and knowledge discovery”, CRC Press, July 29, 2003.

[15] I.T. Jolliffe, “Principal Component Analysis”, Springer Eds., 2002.

[16] K.I. Diamantaras, S.Y. Kung, “Principal component neural networks: theory and applications”, John Wiley & Sons, Inc. New York, NY, USA, 1996.

[17] B. Scholkopf, A. Smola and K.R. Muller, “Kernel principal component analysis”, Advances in Kernel Methods-Support Vector Learning, 1999.

[18] B. Moore, “Principal component analysis in linear systems: Controllability, observability, and model reduction”, IEEE Transactions on Automatic Control, Vol. 26, Issue 1, Feb. 1981.

[19] P Baldi and K Hornik, “Neural networks and principal component analysis: learning from examples without local minima”, Neural Networks, Vol. 2, Issue 1, pp. 53-58, 1989

[20] E. Zio, P. Baraldi, N. Pedroni, “Selecting Features for Nuclear Transients Classification by Means of Genetic Algorithms”, accepted for publication on IEEE Trans. on Nuclear Science, 2006.

[21] G. Gola, E. Zio, P. Baraldi, D. Roverso, M. Hoffmann, “Signal Grouping for Sensor Validation: a Multi-Objective Genetic Algorithm Approach”, HWR-852, OECD Halden Reactor Project, Feb. 2007.

[22] I. Lawrence, L. Kuei, “A Concordance Correlation Coefficient to Evaluate Reproducibility”, Biometrics, Vol. 45, No. 1 (Mar., 1989), pp. 255-268

[23] R. J. Hunt, “Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability”, Journal of Dental Research, 1986, Vol 65, pp. 128-130.

[24] A. Kirschner, M. Hoffmann, “PEANO NNPLS: Advancements in 2002-03”, HWR-741, OECD Halden Reactor Project, April 2004.

[25] R. W. Hamming, “Error-detecting and error-correcting codes”, 1950, Bell System Technical Journal 29(2):pp. 147-160

[26] G.D. Forney Jr, M. Codex, M.A. Mansfield, “On the Hamming Distance Properties of Group Codes”, 1992, IEEE Transactions on Information Theory, Vol. 38(6), pp. 1797-1801.

[27] E. Zio, P. Baraldi, G. Gola, “Feature-based classifier ensembles for diagnosing multiple faults in rotating machinery”, 2007, accepted for publication on Applied Soft-Computing Journal (ASCJ).