# Probabilistic Wavelet Method for Intelligent Prediction of Turbomachinery Damage

Xiaomo Jiang[1], Lijie Yu[2], and Karen Miller[3]

[1,2]*General Electric Company, Power, Global Fleet Services, Data and Analytics, Atlanta, GA 30339, USA*
xiaomo.jiang@ge.com
lijie.yu@ge.com

[3]*General Electric Company, Power, Global Fleet Services, Monitoring and Diagnostics, Atlanta, GA 30339, USA*
karenW.miller@ge.com

## ABSTRACT

This paper develops an innovative integrated methodology for turbomachinery event detection and prediction by using multivariate noisy operation data. The method seamlessly integrates probabilistic method with multiple advanced analytics techniques, including wavelets and entropy information theory. Wavelets based multi-scale principal component analysis is employed to de-noise the raw data for each tag/variable. Probabilistic principal components analysis is further developed to extract useful information from multiple corrected variables, and entropy information feature is extracted as a precursor of the event, the measure of disorder in a thermodynamic system. The proposed method is so-called Wavelet PCA Entropy. The method considers uncertainty in multivariate data, and provides proof-of-concept of advanced analytics for prediction of challenging events in turbomachinery. The feasibility of the presented methodology is demonstrated with the prediction of combustor lean blow out event and data collected from a real-world gas turbine. This study provides a novel intelligent approach to turbomachinery damage diagnostics and prognostics.

## 1. INTRODUCTION

Real-time remote operation monitoring and event prediction of turbomachinery takes advantage of the four critical components: high-speed internet, big data, advanced analytics and domain expertise to provide a state-of-the-art digital solution for asset performance management (APM). It has become a key part of the Internet of Things (IoT) and industry internet ego. The remote monitoring and prediction uniquely integrates interdisciplinary functionality to minimize downtime and improve performance and output more conveniently, easily and effectively than using existing disparate solutions. The integrated solution provides a new world of possibility with advanced APM capabilities, such as cloud-based monitoring & diagnostics, asset lifecycle management, predictive maintenance, and operation intelligence.

In the marketplace of remote monitoring & diagnostics (M&D), the question arises when is the anomaly expected to happen on the monitored assets and how much longer can the user operate the machine with this anomaly. The conventional analytics based on classical statistical techniques are hardly able to answer these questions. The availability of huge volume of machinery operation data provides the unprecedented opportunity for advanced analytics in the remote monitoring domain. This paper provides advanced data driven analytics for complicated event monitoring and prediction, to enable integration of a wide range of data, including fleet knowledge and unit operational data into the platform. It also offers unified view of all predictive or anomaly alerts/alarms/advisories in a single pane of glass. Single unified, integrated data storage for all available data allows all data to be utilized and analyzed seamlessly together as an integrated whole instead of having silo'd, non-integrated, non-interoperable data stores attached to different applications in the back end, as the traditional M&D services do.

This paper presents an innovative method for anomaly prediction by seamlessly integrating probabilistic method with multiple advanced analytics techniques, including multiresolution wavelet analysis and entropy information theory. Wavelets based multi-scale principal component analysis is employed to de-noise the raw data for each tag/variable. Probabilistic principal components analysis (PPCA) is further developed to extract useful information

[1]Corresponding author: 1-678-871-9281 (Tel).

from multiple corrected variables, and entropy information feature is extracted as a precursor of the event, the measure of disorder in a thermodynamic system. The proposed method is so-called Wavelet PCA Entropy (WAPE), which considers uncertainty in multivariate data, and provides proof-of-concept of advanced analytics for prediction of challenging events in turbomachinery. The feasibility of the presented methodology is demonstrated with the prediction of combustor lean blow out event and data collected from a real-world gas turbine.

## 2. METHODOLOGY

The proposed anomaly detection method includes three advanced techniques: multiresolution wavelet decomposition, probabilistic PCA, and entropy information feature, as described below.

### 2.1. Wavelet Signal Processing

Wavelets consist of a family of mathematical functions used to represent a signal in both time and frequency domains. A *wavelet transform* decomposes the signal into two sub-signals: *approximation* (the average of adjacent elements) and *details* (difference of adjacent elements). The *approximations* represent the high-scale, low-frequency components of the signal, while the *details* represent the low-scale, high-frequency components. As such, wavelets provide an effective and efficient approach to obtain a multi-resolution representation of a signal. This representation provides a hierarchical framework for interpreting the information context in a signal. At different resolutions, the details of a signal characterize different physical structures of the scene. At a fine resolution, these details correspond to the transient changes which provide the signal "context". For a given wavelet basis and decomposition level, the wavelet transform of a signal has been demonstrated to be unique and invariant [1-3]. Coifman and Wickerhauser [4] proposed the *wavelet packet transform* (WPT) analysis to allow for a finer and adjustable resolution in the high frequencies (details). Compared with conventional wavelet transform methods, the wavelet packet transform method is a more effective approach to extract features from either stationary or non-stationary signals to represent the underlying dynamic systems [3,5].

In the DWT analysis, the wavelets, $\psi_{j,k}(t)$, are obtained from the basis function (also known as *mother* or generating wavelet) $\psi(t)$ by simple scaling and translation in the dyadic form as follows [1]:

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k),\ j,k \in Z,\ \psi \in L^2(R) \quad (1)$$

where *t* represents a continuous time variable, $k$ and $j$ denote the time and frequency indices, respectively, and *Z* is the set of all integers. The notation $L^2(R)$ represents the square summable real number space.

### 2.2. Multiscale Wavelet PCA Denoising

With the multiresolution wavelet coefficients obtained from DWT analysis, the multiscale principal components analysis (MPCA) is developed to perform signal denoising on each tag of the problem under investigation. The MPCA approach generalizes the usual principal components analysis (PCA) of the multivariate signals as a matrix by performing simultaneously a PCA on the different levels of DWT coefficients of a signal. In addition, a PCA is performed also on the final reconstructed matrix for denoising purpose. Recently the multiscale wavelet PCA (MWPCA) has been demonstrated to be effective in data cleansing [6-8].

PCA technique is usually used to reduce a higher-dimensional data set to lower dimensions for subsequent analysis. It requires computation of the eigenvalue decomposition or singular value decomposition of a data set, usually after mean centering the data for each attribute. The aim of MWPCA approach is to remove the noise in the original signal through selecting the principal components from the wavelet signal decomposition. A thresholding rule is employed in this method to remove potential noises from the decomposed coefficients, where the components associated with eigenvalues greater 0.05 times the sum of all eigenvalues.

The denoising performance is then assessed quantitatively by the commonly used SNR measure (Signal to noise ratio), expressed as

$$SNR = 10 \cdot log \frac{\sum_i f^2(t_i)}{\sum_i [f(t_i) - \hat{f}(t_i)]^2} \quad (2)$$

where $\hat{f}$ is the cleaned signal. The summation is performed over the signal length and *log* is the logarithm base 10.

The Welch method [9] is further employed to evaluate the denoising data. The method produces the spectral density from a finite-length signal by averaging the periodograms of overlapped, windowed signal sections. The periodogram is obtained by a short time Fourier transform. The average enables to reduce the variance or noise in the estimated power spectra, which is different from other existing methods such as the multiple signal classification method or the eigenvector method. The denoising performance will be further assessed by the probabilistic PCA dimension reduction presented subsequently.

### 2.3. Probabilistic Principal Components Analysis

After the multivariate time series data are cleaned, the probabilistic principal component analysis (PPCA) approach is developed to (1) reduce data dimensionality, (2) address the multivariate correlation, and (3) consider data uncertainty. Principal component analysis (PCA) [10] is a well-established statistical method for dimensionality reduction and has been widely applied in data compression, image processing, exploratory data analysis, pattern

recognition, and time series prediction [11]. PCA involves a matrix analysis technique called eigenvalue decomposition. The decomposition produces eigenvalues and eigenvectors representing the amount of variation accounted for by the principal component and the weights for the original variables, respectively. Its main objective is to transform a set of correlated high dimensional variables to a set of uncorrelated lower dimensional variables (principal components). The important property of PCA is that the principal component projection minimizes the squared reconstruction error in dimensionality reduction. However, the PCA is not based on a probabilistic model such that it cannot handle data uncertainty. The probabilistic principal component analysis method proposed by Tipping and Bishop [12] is employed in this study to address this issue.

The PPCA is derived from a Gaussian latent variable model which is closely related to statistical factor analysis. The factor analysis is a mathematical technique widely used to reduce the number of variables (dimensionality reduction), while identifying the underlying factors that explain the correlations among multiple variables [13, 14]. For the convenience of formulation, let $\mathbf{y}_i = \tilde{f}(t_i) \in \Re^q$ represent the $q$-dimension real numbers, and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]^T$ denote the $N \times q$ data matrix representing the $q$ variables, each containing $N$ cleaned time series data points, $\tilde{f}(t_i)$. Let $\mathbf{\Phi} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N]^T$ be the $N \times d$ data matrix with $\boldsymbol{\theta}_i \in \Re^d$ ($d \le q$) representing $d$ latent variables (factors) that cannot be observed, each containing the corresponding $N$ positions in the latent space. The latent variable model relates the correlated data matrix $\mathbf{Y}$ to the corresponding uncorrelated latent variable matrix $\mathbf{\Phi}$, expressed as

$$\mathbf{y}_i = \mathbf{W}\boldsymbol{\theta}_i + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i, \; i = 1, 2, \ldots, N \tag{3}$$

where the $q \times d$ weight matrix $\mathbf{W}$ describes the relationship between the two sets of variables $\mathbf{y}_i$ and $\boldsymbol{\theta}_i$, the parameter vector $\boldsymbol{\mu}$ consists of $q$ mean values obtained from the data matrix $\mathbf{Y}$, i.e. $\boldsymbol{\mu} = (1/N)\sum_{i=1}^{N} \mathbf{y}_i$, and the $q$-dimensional vector $\boldsymbol{\varepsilon}_i$ represents the possible error or noise in each variable $\mathbf{y}_i$, which may not be completely removed by MWPCA approach. The error $\boldsymbol{\varepsilon}_i$ is usually assumed to consist of independently distributed Gaussian variables with zero mean and unknown variance $\boldsymbol{\psi}$, same as we did in the previous section.

The probabilistic principal component analysis was derived from the statistical factor analysis, with an isotropic noise covariance $\sigma^2\mathbf{I}$ assumed for the variance $\boldsymbol{\psi}$. In particular, it was shown that, with the Gaussian distribution assumption for the latent variables, the maximum likelihood estimator for $\mathbf{W}$ spans the principal subspace of the data even when the $\sigma^2$ is non-zero. The use of the isotropic noise model $\sigma^2\mathbf{I}$ makes the probabilistic PCA technically distinct from the classical factor analysis. The former is covariant under rotation of the original data axes, while the latter is covariant under component-wise rescaling. In addition, the principal axes in the PPCA are in the incremental order, which cannot be realized by the factor analysis. Refer to [12] for details of derivative of PPCA.

### 2.4. Entropy Feature Extraction

The negative of the logarithm of the probabilistic distribution of an event is often used to measure the information of an event because the logarithm is additive for independent variables. The information amount forms a random variable whose expected value, or average, is called as *entropy* or more specifically Shannon entropy introduced by Claude Shannon in 1948 [15]. In thermodynamics, **entropy** is commonly understood as a measure of *disorder* or *unpredictability*. It measures the number of specific ways in which a thermodynamic system may be arranged. According to the second law of thermodynamics, the entropy of an isolated system never decreases. Entropy is extracted as a feature from the principal components obtained from PPCA, and used to detect the possible change on *disorder* of the thermodynamic system, e.g., gas turbine or its main component. Note that entropy only considers the probability of observing a specific event, so the information it encapsulates is referred to that about the underlying probability distribution, not the meaning of the events themselves.

In practice, a mechanical system such as gas turbine or its combustion is not completely an "isolated" system. The entropy obtained from the operational time series data of a system is expected to vary over the time, i.e., variability. However, a significant consistent outlier may be an indicator of anomaly for the system under operation.

In this study, the non-normalized Shannon Entropy is developed as the feature for anomaly detection. The entropy is defined as

$$Ent(k) = -\sum_{j=1,S}[PC_{k,j}^2 \log(PC_{k,j}^2)] \tag{4}$$

The entropy is calculated at a continuously rolling window with the given size of one day data at 5 min interval (288 points) in this study. The obtained entropy is used as the feature to evaluate the status of a system using anomaly detection strategy as explained subsequently in the application procedure.

## 3. APPLICATION PROCEDURE

Figure 1 shows the generalized procedure and strategy to

implement the proposed WAPER methodology for damage detection, as explained below as 16 steps (denoted by numbers in Fig. 1):

1) Read raw time series data acquired from multiple sensors installed on a gas turbine. Table 1 shows the example sensors used to acquire operational data. The sensor data includes various operation variables, such as compressor inlet and exhaust temperature, pressure, and generator output, as well as different bands of combustion dynamics sensors, such as CDAL and CDAM. These sensors measure the operational performance of the turbine and dynamics of the combustor, generally in the form of time series. The data at regular 5-minute interval over a 3-month period of calendar time is employed in this study. The multiple time series data are obtained for q variables, each having M data points, yielding a raw data matrix Rq×M for signal processing and damage diagnosis. The 5-min interval data is used in the example presented in this paper for demonstration purpose.

2) Perform data validation on the acquired time series data for the unit under investigation. Techniques may include graphical plots, outlier analysis and data filtration. Outliers tend to pull the mean value towards themselves and inflate the variance in their direction. Therefore, the outliers will largely affect the moment characteristics of the data. Outliers may be identified through graphical plots of each raw data set. These outliers should be removed for further analysis only with proper justification. Some sensed data points may be inconsistent with the expectation of the majority elements of the series. These data points are usually referred to be outliers. These outliers may result from measurement errors and anomaly, which cannot be used to represent the normal operational condition of that unit. In this study, data filtration is performance to ensure that the sensor data represents the unit under normal operation.

3) Build the data set with a rolling window for subsequent signal processing at each time step. The sensor data is divided into two groups: operation and combustion dynamics variables. One day is selected as the rolling window size. The rolling window strategy enables the proposed methodology applicable to continuously evaluate the status of the machine under monitoring.

4) Perform discrete wavelet packet transform of the i-th time series $\mathbf{S}_{i\times N}$ ($i$ = 1, …, $q$), each containing $N$ preprocessed data points. The DWPT approach decomposes the raw time series into multi-resolution time-frequency domains for each variable. Each time series will be decomposed on $p$ sets of coefficients $\mathbf{A}_{p\times N}$, in which $p = 2^j$ is the number of wavelet coefficients for $j$-th level decomposition, e.g., $j$ = 1, 2, and 3 for the 3-level decomposition.

5) Perform MWPCA on each data set. The integrated wavelet PCA signal processing captures subtle details of signals meanwhile assessing noise, error and incoherence context in a direct means, thus avoiding the conventional under-denoising and over-denoising issues.

6) Reconstruct each time series signal from the cleansed wavelet coefficients by using Eq. (2).

7) Assess the effectiveness of the MWPCA denoising approach quantitatively by using the SNR measure and graphically by using spectral analysis.

8) Determine the reduced dimension $d$ using PCA approach and a predefined threshold for all variables (95% used in this study). Note that the threshold used in this step is to perform data fusion to combine useful information from multiple sensor data.

Table 1. Example of tags for raw data

| Tag | Description | Units | Group |
|---|---|---|---|
| DWATT | Generator Power | MW | Operation |
| AFPAP | Ambient Pressure | in Hg | Operation |
| CDAB_1 | Combustion Dynamics Amplitude Blowout Band Can 1 | psi | Comb Dynamics |
| CDAL_2 | Combustion Dynamics Amplitude Low Band Can 2 | psi | Comb Dynamics |
| CDAL_3 | Combustion Dynamics Amplitude Low Band Can 3 | psi | Comb Dynamics |
| CDAM_3 | Combustion Dynamics Amplitude Mid Band Can 3 | psi | Comb Dynamics |
| CDAM_4 | Combustion Dynamics Amplitude Mid Band Can 4 | psi | Comb Dynamics |
| CDAM_5 | Combustion Dynamics Amplitude Mid Band Can 5 | psi | Comb Dynamics |
| CTD | Compressor Discharge Temperature | deg F | Operation |
| CTIM | Compressor Inlet Temperature | deg F | Operation |
| FTG | Gas Fuel Temperature | deg F | Operation |

9) Produce the reduced data matrix, $\Phi_{d\times N}$, using probabilistic PCA approach. The multivariate statistical analysis method is employed to fuse multivariate data, considering data uncertainty and correlation. The method incorporates statistical factor analysis, matrix Eigenvalue decomposition, and maximum likelihood estimation. Its purpose is to integrate useful information from multiple sensors, considering uncertainty and correlation in the multivariate sensor data.

10) Calculate entropy value of both operation and combustion dynamics groups based on their reduced principal components at each time step.

11) Calculate the ratio of entropy between combustion dynamics and operation groups given a time step.

12) Calculate the sigma of the entropy ratio feature at the validation domain which represents healthy status of the machine. The sigma is used as the baseline for upcoming

operation.

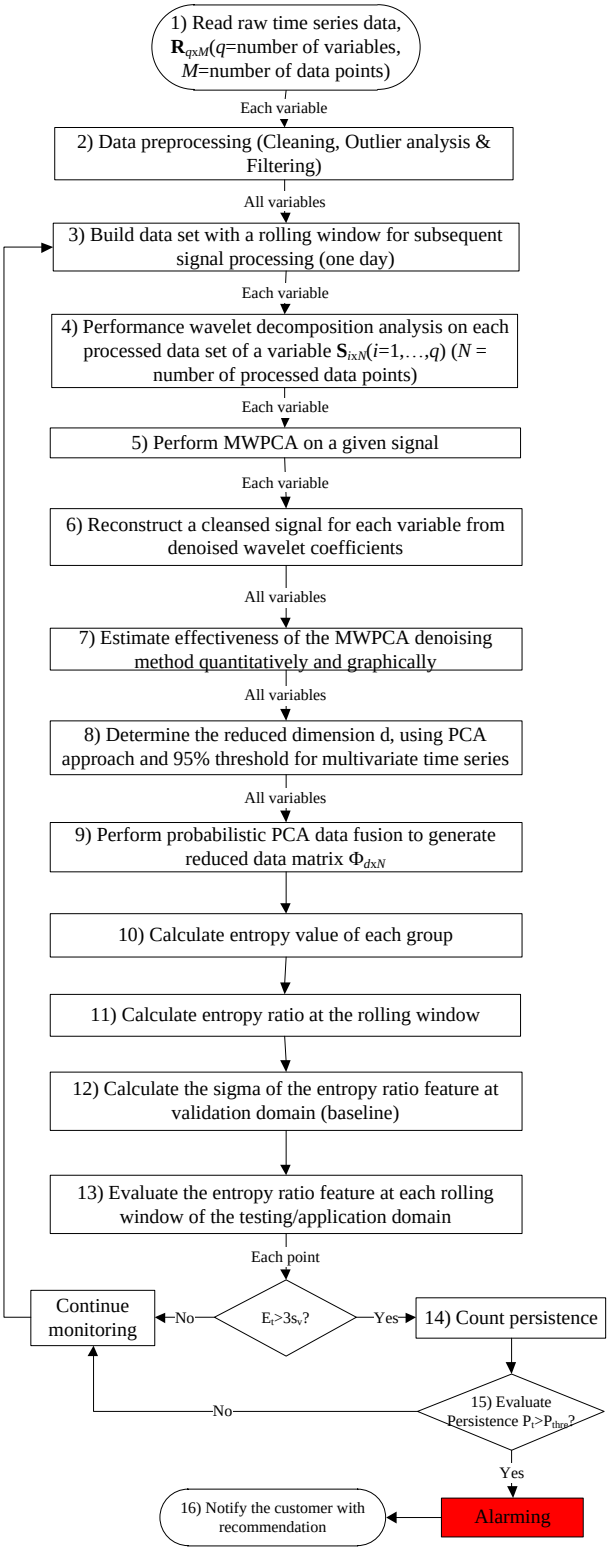


Figure 1. Flow chart of methodology implementation

13) Evaluate the entropy ratio at each rolling window at the testing or application domain. The judgement will be made based on whether the calculated entropy ratio exceeds $3\sigma_{val}$, which is obtained from the validation data set at previous step. If the condition is met, go to next step. Otherwise, keep monitoring the machine.

14) Count the times of the calculated entropy ratio exceeding the $3\sigma_{val}$.

15) Evaluate whether the accumulated persistence count exceeds the reset threshold, e.g., 5 alerts in 3 hours. If the persistence condition is violated, an alarm will be triggered.

16) Provide actionable maintenance recommendation to notify the customer if the alarm is confirmed with diagnostics results. The diagnosis results will not only facilitate the decision-making in alarming logics development and condition evaluation for the equipment, but also assist the asset management of the equipment by scientific schedule of inspection and maintenance and effective management of parts procurement.

## 4. ILLUSTRATIVE EXAMPLES

The proposed methodology and process is demonstrated with a set of operational data and a combustor lean blowout (LBO) trip event obtained from a real heavy duty gas turbine. Per the inputs of subject matter experts, the time series data of the 28 combustion dynamics variables and 5 operation related main variables was collected at 5-min interval from Jan 1st to Mar 27th, right before the LBO trip on Mar 27th, resulting in about $M$ = 26000 raw data points in each variable. The example is used to demonstrate the effectiveness of the proposed probabilistic signal processing methodology and process shown in Fig. 1 on event prediction.

### 4.1. Problem Background

Usually the combustor will be operated over a wide range of operating conditions with a higher level of efficiency. The blowout becomes a main concern in the combustor of both military and commercial aircrafts, or even the industrial heavy-duty gas turbine, during sudden changes in throttle setting. The blowout frequently occurs when the air flow change doesn't catch up with the fuel flow variation due to the change of machine operation mode, such as quick deceleration. LBO is extremely detrimental for both industrial gas turbines and aero-engines. In the former, LBO leads to prolonged shutdown and relighting involving productivity loss. For the latter during throttling operation, the fuel flow is suddenly reduced. Due to the inertia of the compressor, reduction in airflow takes place at a much slower rate. The consequent sudden decrease in equivalence ratio can lead to LBO on aircraft engines, which can have fatal consequences.

A priori determination of the LBO margin is difficult because it is dynamically altered in the presence of thermos-acoustic instabilities. This calls for development of strategies for early detection of imminent blowout and adoption of appropriate

measures to mitigate it. Although the industrial gas turbines mostly operate in a lean premixed mode, in aircraft engines, the fuel is admitted close to the burner, leading to partially premixed combustion. The LBO problem is employed in this study as an example to demonstrate the effectiveness of the proposed advanced analytics methodology.

### 4.2. Data

All operational data measured from an operating machine unavoidably contain errors, outliers, missing values, and noise, in addition to the variation of time series. As an example, Figure 2 shows the raw time series data for the generator output, turbine speed (%), and one of the combustion dynamics amplitude blowout bands (CDAB). Note that the data gap from Feb 25$^{th}$ to Mar 5$^{th}$ indicates that the unit is off. Several observations can be drawn from the time series plots of the raw data. First, it is difficult to visually identify any significant change from the raw data series over the time, indicating that advanced analytics techniques are needed for signal processing of the data to develop automatic anomaly detection algorithms. The CDAB variable (the 3rd time series in Fig. 2) shows a gradual decreasing trend over time approaching to the trip, but the variable quantity is still under the operation boundary. The pattern change would be an indicative to the LBO event. However, it is unclear whether this trend is reasonably correlated to other sensor variables or not. Multivariate analysis techniques are therefore needed to facilitate the decision making on the anomaly diagnostics and prognostics of a complicated system with multivariate time series data.

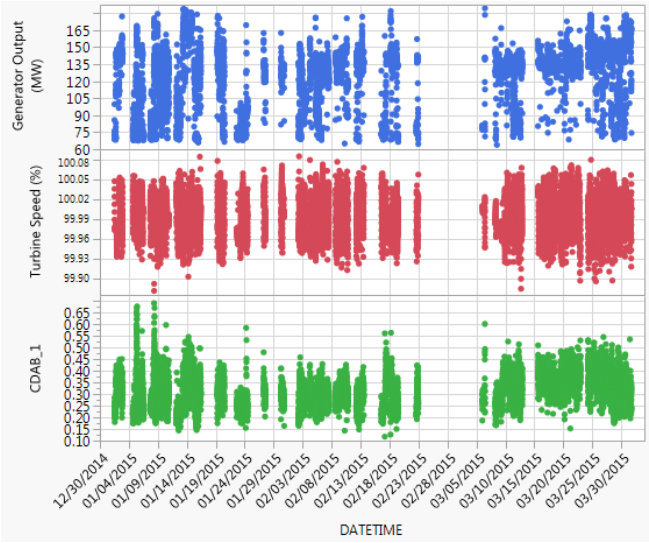


Figure 2. Sample data used in this example

### 4.3. Data Processing

After removing the obvious outliers or unavailable points from each time history data, the same number of data points are taken for all variables, leading to a data matrix with the same dimension $q \times N$, where $q = 47$ and $N = 25900$ (in comparison to $N = 26000$ in the raw data set). In addition, the multivariate time series have different magnitude or units. In order to use multiple time series variables simultaneously for damage diagnosis and prevent the undue domination of variables with large numerical values over the variables with small numerical values, each time series is divided by its corresponding maximal value, thus normalizing the variables into dimensionless vector with the same range of -1 to 1. The normalized multivariate data series will be used in the subsequent analyses.

### 4.4. Multiscale Wavelet PCA Denoising

It is found that three decomposition levels of DWT using the Daubechies wavelet of order 8 described previously can adequately characterize the details of each time series data in this example. Given a time series, the multiscale wavelet PCA method described previously is first applied to each measured time series. A cleaned time series is then reconstructed from the denoised wavelet coefficients. For the 3$^{rd}$ combustion dynamics amplitude low band (CDAL3) data, for example, the signal-to-noise ratio, $SNR = 6.98$, is obtained. Figure 3 shows the comparison of raw and denoised data for CDAL3 sensor. It is observed that main information in the signal has been reserved. Furthermore, the power spectrum density (PSD) is employed as a graphical metric to indicate the denoised data. Figure 4 shows the PSD obtained by using the Welch method for the cleansed data. It is clearly observed that the PSD characterizes the high-frequency components in the cleaned data. Obviously the high-frequency characteristics of the signal makes the denoising of the time series data challenging.

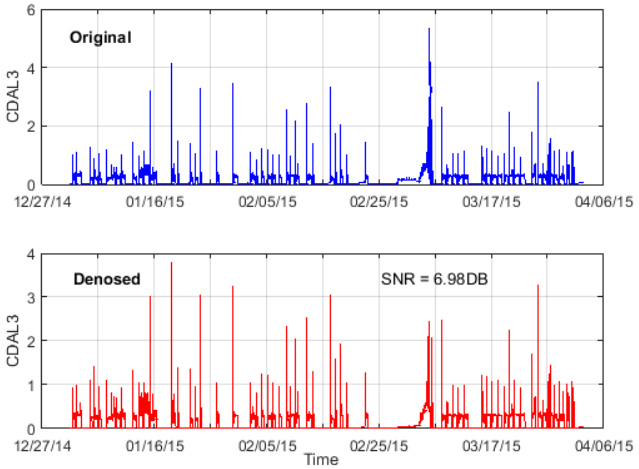


Figure 3. Example of raw and denoised data (SNR = 6.98DB)

### 4.5. Probabilistic PCA

Before applying the probabilistic PCA to reduce the dimensionality and quantify the data uncertainty, the standard PCA analysis is applied to determine the proper number of the reduced dimensionality. The number of principal

components is determined based on the 33 × 25900 cleansed data matrix **Y** by predefining at least 95% information in the original data to be considered. The value of $d = 2$ is obtained for the cleaned data matrix, which accurately accounts for 98.3% information in the original data, while $d = 3$ is obtained for the raw data matrix, which accounts for 96.3% information in the original data. Figure 5 shows the accumulation of information in the original data as the principal components increases. Obviously the smaller principal components are required to represent the information in the original data if the cleaned data is analyzed.

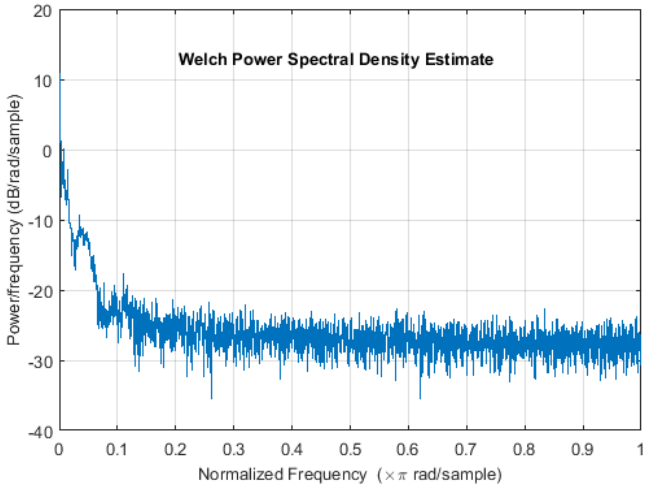


Figure 4. Welch power spectral density estimate of CDAL3

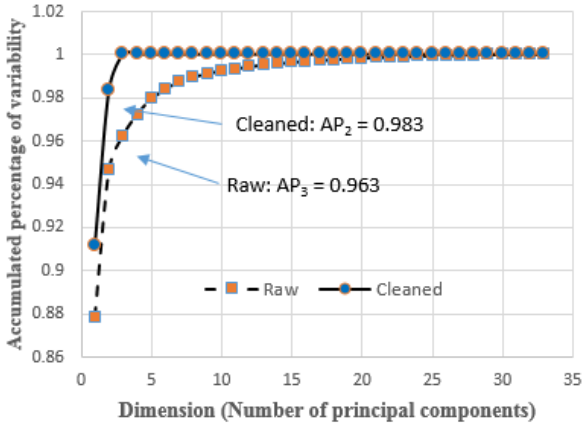


Figure 5. PCA dimension reduction results of raw and cleaned data.

For each of the raw and cleaned 33 × 25900 data matrix, the maximum likelihood estimates for the PPCA variability parameters $\sigma^2$ and **W** are obtained. Table 1 summarizes part of the coefficient matrix **W** corresponding to the first two principal components of the cleaned and raw data, respectively. Each cell shows the weight of the variable contributing to the corresponding principal component. For example, the value of 0.864 in the 1st column and the 2nd row represents the weight of the compressor discharge pressure contributing to the first principal component. It is observed that five operation variables make a major contribution (its weight is greater than 0.04) to the first principal component (marked in shadow cells in the first column), while the combustion dynamics variables make a major contribution to the second principal component. The coefficient matrix demonstrates that it can effectively identify the critical variables which make significant contribution to the principal components. Note that the information can also be utilized to improve the accuracy of failure predictive modeling in future research.

Table 2 PPCA coefficients for two PCs of cleaned data

| Variables | PC1(91.1%) | PC2(7.2%) |
|---|---|---|
| AFPAP | 0.04780 | 0.00007 |
| CTD | 0.86398 | 0.00120 |
| DWATT | 0.13753 | 0.00019 |
| CTIM | 0.09989 | 0.00014 |
| FTG | 0.47154 | 0.00066 |
| CDAB_1 | 0.00051 | -0.03622 |
| CDAB_2 | 0.00037 | -0.02631 |
| CDAB_3 | 0.00039 | -0.02737 |
| CDAB_4 | 0.00044 | -0.03106 |
| CDAB_5 | 0.00043 | -0.03060 |
| CDAB_6 | 0.00045 | -0.03178 |
| CDAB_7 | 0.00043 | -0.03044 |
| CDAB_8 | 0.00043 | -0.03026 |
| CDAB_9 | 0.00044 | -0.03103 |
| CDAB_10 | 0.00049 | -0.03489 |

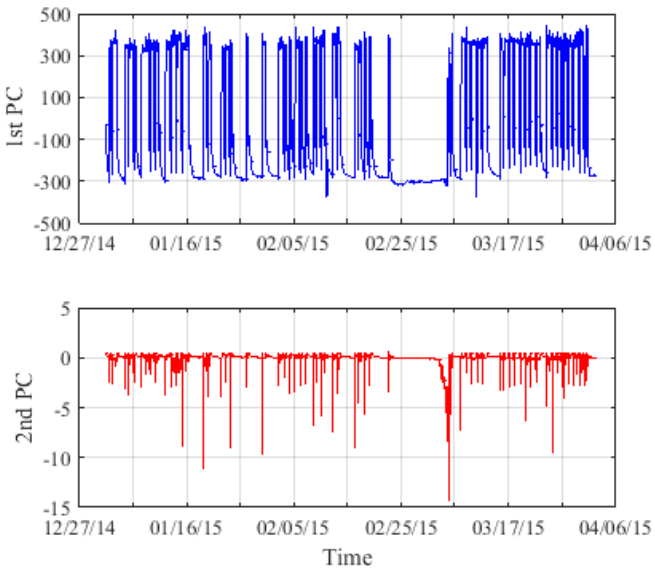


Figure 6. Principal components of cleaned data

The resulting coefficients $\mathbf{W}_{ML}$ are used for the data matrix to produce the variance matrix $\boldsymbol{\Sigma}_{ML}^{-1}$ and the $d \times N$ reduced data

matrices $\mathbf{\Phi}^*$. Figure 6 shows the cleansed in terms of the first and second principal components. It is difficult to identify any historical trend pattern change even from both the first and second principal components (totally accounting for 98.3% information in the original data) obtained from the cleansed data. This implies that further feature extraction is needed from the principal component for damage diagnosis.

### 4.6. Entropy Feature Extraction

Entropy is calculated by using Eq. (4) for each rolling window size (288 point). Figure 7 shows the obtained 3-month entropy series. Two observations can be obtained from the entropy series. First, four sparks are observed over the 3-month data. The last spark close to the LBO trip indicates significantly negative larger than others. This may be an indicator of the possible anomaly. Second, the entropy decreasing trend (negatively increasing) over the operation time approaching to the event is observed, represented by the linear regression trend line. Both observations provide possible indicators to identify the event.

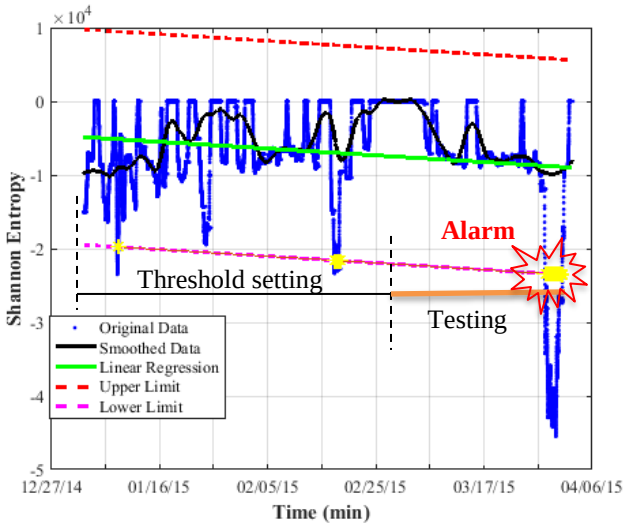


Figure 7. Anomaly alerting from entropy feature data

### 4.7. Anomaly Detection

The first 15000 entropy data are assumed to be healthy without any LBO event, and used to calculate the standard variation of the entropy series. The alert may be triggered if the actual entropy exceeds the 3 times standard deviation. Since the data points may be outliers, a persistence logic is employed to trigger a real alarm. As described in the procedure, in this study an alarm will be triggered if the accumulated alerts exceed 5 within 3 hours.

This anomaly detection logic is applied for the rest of data. As shown in Fig. 7, an alarm is triggered on early Mar 27th, about 4 hours in advance of the trip.

## 5. CONCLUDING REMARKS

This paper presents an innovative probabilistic signal processing methodology for anomaly prediction of high-fidelity turbomachine equipment or components. The methodology and procedure seamless integrates the wavelet multi-resolution decomposition, probabilistic principal component analysis, and entropy information to provide anomaly prediction in a turbomachine, considering possible uncertainty and noise in the sensed multivariate time historical data. The discrete wavelet transform analysis is employed to decompose a time series signal into different levels of wavelet coefficients. These coefficients represent the multiple time-frequency resolutions of a signal. Multiscale principal component analysis is then applied to wavelet coefficients to remove possible imperfections.

Furthermore, the probabilistic principal component analysis approach is developed to reduce dimensionality and to address multivariate correlation and data uncertainty for damage diagnostics. The Shannon entropy is calculated at a rolling window of the time series and used as the indicator for the anomaly detection. A generalized framework and process is developed to implement the proposed probabilistic signal processing methodology. The proposed method and process is demonstrated with a set of 47-variable sensor data collected from a real-world gas turbine with a LBO trip event.

Numerical results show that 1) the anomaly cannot be readily identified via the multivariate time series analysis of raw data with uncertainty and noise, implying that the multivariate signal processing plays a key role on accurately identifying the possible damage in the turbomachinery, and 2) the proposed methodology and strategy provide a promising and powerful tool in addressing imperfections in multiple time historical data and facilitating the decision-making on the default diagnostics using multivariate data, taking into account uncertainty and data correlation.

The proposed methodology provides an advanced state-of-the-art approach and fundamental to process multivariate time series data, to develop damage diagnostics strategy to evaluate the turbomachine condition, thus facilitating cost-effective schedule for timely preventive maintenance, and ensuring product safety and increasing product availability and customer satisfaction. The multivariate signal processing method presented in this paper has potential to largely enhance diagnostics accuracy and facilitate prognostics of complicated turbomachine systems. It can be extended to investigate more complicated situations in turbomachine dynamic systems, such as huge amount of monitoring data sets (big data), missing data in measurements and multiple-multivariate measurements with different boundary conditions.

## Reference

[1]. Daubechies, I. (1988), Orthonormal Bases of compactly supported wavelets, *Communication on Pure and Applied Mathematics* 41(7): 909–996. DOI: 10.1002/cpa.3160410705.

[2]. Daubechies, I., *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics: Pennsylvania, 1992. doi.org/10.1137/1.9781611970104.

[3]. Mallat, S. (1989), A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(7):674–693. DOI: 10.1109/34.192463.

[4]. Coifman, R.R., Wickerhauser, M.V. (1992), Entropy-based algorithms for best basis selection, *IEEE Transaction on Information Theory* 38(2): 713–718. DOI: 10.1109/18.119732.

[5]. Jiang, X., Mahadevan, S., Adeli, H. (2007), Bayesian wavelet packet denoising for structural system identification, *Structural Control and Health Monitoring* 14(2): 333–356. DOI: 10.1002/stc.161.

[6]. Aminghafari, M., Cheze, N., Poggi, J.M. (2006), Multivariate de-noising using wavelets and principal component analysis, *Computational Statistics & Data Analysis*, 50(9): 2381–2398. doi.org/10.1016/j.csda.2004.12.010.

[7]. Mostacci,E., Truntzer, C., Cardot, H., Ducoroy, P. (2010), Multivariate denoising methods combining wavelets and principal component analysis for mass spectrometry data, *Proteomics* 10(14): 2564-2572. doi: 10.1002/pmic.200900185.

[8]. Vijaykumar, D.S., Patil, C.G., Ruikar, S.D. (2012), Wavelet based multi-scale principal component analysis for speech enhancement, *International Journal of Engineering Trends and Technology*, 3(3): 397-400.

[9]. Stoica,P., Moses, R.L., *Introduction to Spectral Analysis*. New Jersey: Prentice-Hall, Englewood Cliffs, 1997.

[10]. Hotelling, H. (1933), Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* 24(6): 417–441. doi.org/10.1037/h0071325.

[11]. Joliffe, I.T., *Principal Component Analysis*, Springer, New York, 2002.

[12]. Tipping, M.E., Bishop, C.M. (1999), Probabilistic principal component analysis, *Journal of the Royal Statistical Society*: Series B (Statistical Methodology) 61(3): 611–622.

[13]. Cooley, W.W., Lohnes, P.R., *Multivariate Data Analysis*, Wiley & Sons, New York, 1971.

[14]. Tukey, J.W., Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts, 1977.

[15]. Shannon, C.E. (1948), A mathematical theory of communication, *Bell System Technical Journal* 27(3): 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

## BIOGRAPHIES

**Xiaomo Jiang** received his Ph.D. in Civil Engineering from The Ohio State University in 2005, Post-doctoral Researcher from Vanderbilt University in 2007. Currently he is a Senior Manager of Data & Analytics at GE, leading advanced analytics development for monitoring, diagnostics, and prognostics of thousands of different OEM turbines, as well as predictivity analytics and their industrial internet applications on energy assets. Dr. Jiang has authored 1 book, 4 chapters, and over 60 research articles in fields of engineering, computing, and statistics. He is a recipient of multiple patents and awards, and has been invited to serve as a peer reviewer, scientific committee, editorial board or associate editor for over 40 different international journals and conferences.