Adapting nearest neighbors-based monitoring methods to irregularly sampled measurements

Inês M. Cecílio[1], James R. Ottewill[2], and Nina F. Thornhill[3]

[1,3] *Centre for Process System Engineering, Department of Chemical Engineering, Imperial College London, London SW7 2AZ, UK*
i.cecilio09@imperial.ac.uk
n.thornhill@imperial.ac.uk

[2] *ABB Corporate Research Center, ul. Starowiślna 13a, 31-038 Kraków, Poland*
james.ottewill@pl.abb.com

ABSTRACT

Irregularly spaced measurements are a common quality problem in real data and preclude the use of several feature extraction methods, which were developed for measurements with constant sampling intervals. Feature extraction methods based on nearest neighbors of embedded vectors are an example of such methods. This paper proposes the use of a time-based construction of embedded vectors and a weighted similarity metric within nearest neighbor-based methods in order to extend their applicability to irregularly sampled measurements. The proposed idea is demonstrated within a method of univariate detection of transient or spiky disturbances. The result obtained with an irregularly sampled measurement is benchmarked by the original regularly sampled measurement. Although the method was originally implemented for off-line analysis, the paper also discusses modifications to enable its on-line implementation.

1. INTRODUCTION

One of the early steps in the PHM architecture is feature extraction from raw sensor data. Feature extraction aims to retain only the information that is relevant for classification and diagnostics, thus reducing the dimensionality of the raw data space and the risk of misclassification (Russell et al., 2000). Examples of features extracted for classification and diagnostics include the Hotelling T-square statistics, wavelet coefficients, non-linearity of the time series (Thornhill, 2005), and occurence of spiky disturbances (Cecílio et al., 2014).

A common challenge to carry out feature extraction in real systems is is the quality of the measurements available. A usual quality problem is that the interval between samples in a measurement is not constant. Figure 1 shows an example of such irregularity in a measurement from a real gas processing plant. This irregularity may arise for instance from problems with data communication. Another cause is data compression, which is done after sampling in order to save memory. Compression is done either by eliminating samples or by substituting the values of the samples by a constant value, for example, the average over a period.

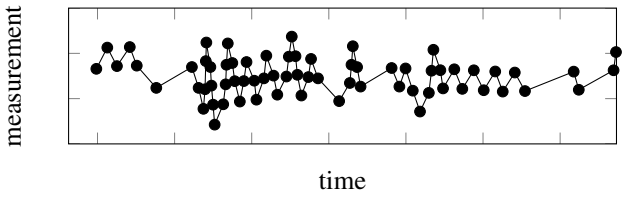


Figure 1. Example of a measurement with irregular sampling intervals obtained from a real gas processing plant.

Several methods for feature extraction were developed for measurements with constant sampling intervals Δt and are not applicable to measurements such as those in Figure 1. For instance, several methods obtain the spectral information of the measurements from their Fourier transforms and wavelet decomposition. However, both techniques assume that the measurement samples are taken at regular intervals. Applications of these in the process monitoring were given by Thornhill et al. (2002); Choudhury et al. (2004); Tangirala et al. (2007); Zang & Howell (2007); Babji & Tangirala (2010). Methods that use cross-correlation, for example to extract time lags (Bauer & Thornhill, 2008), also require regularly-sampled data.

This paper focuses on feature extraction methods that use nearest neighbors of embedded vectors. Embedded vectors are segments of a time series. Nearest neighbors are the segments from a time series which are most similar to a ref-

erence segment (Chandola et al., 2009). Nearest neighbor-based methods have been used successfully to extract the non-linearity of a time series (Thornhill, 2005), time lags (Stockmann et al., 2012), and the occurence of transient or spiky disturbances (Cecílio et al., 2014). However, none of these methods is directly applicable to irregularly sampled measurements. The reason is that the nearest neighbors approach implies measuring the similarity between embedded vectors. The conventional similarity measures are defined between two ordered sequences $\mathbf{p}$ and $\mathbf{q}$ which have the same number m of samples and whose samples are synchronized. For example, the Euclidean distance metric which is used in the references mentioned above is defined as

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{m}(p_i - q_i)^2}\,. \tag{1}$$

However, the segments represented by the embedded vectors normally span a constant interval of time. Therefore, in irregularly sampled measurements those segments will have a varying number of samples and varying intervals between samples. Hence, the conventional similarity measures are not directly applicable.

The contribution of this paper is to reformulate the construction of embedded vectors and the computation of similarity for the case of irregularly sampled measurements. As a result, the new formulation extends the applicability of methods based on nearest neighbors of embedded vectors.

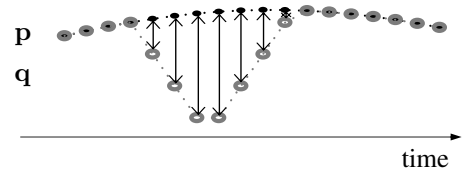
The paper is structured as follows. Section 2 provides background on the analysis of irregularly sampled time series and time-based construction of embedded vectors. Section 3 explains the proposed techniques to construct embedded vectors and to compute similarity in the case of a measurement with irregular sampling rate. These techniques are applicable in the context of any nearest neighbor-based method. For brevity, section 4 demonstrates the techniques in one particular method, which detects and identifies transient disturbances (Cecílio et al., 2014). The demonstration uses the same case study as Cecílio et al. (2014) in order to have a benchmark for the results. Section 5 closes with conclusions.

2. BACKGROUND

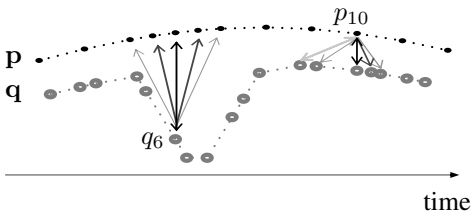
2.1. Analysis of irregularly sampled time series

Research in irregularly sampled time series is commonly found in domains such as astronomy (Scargle, 1989; Bos et al., 2002), finance (Zumbach & Müller, 2001), and geophysics (Rehfeld et al., 2011).

Weighting methods are one of the type of methods to analyse irregularly sampled time series. They generalize measures, such as distance and correlation, which are conventionally defined for pairs of aligned samples (Rehfeld et al., 2011). This paper uses a weighting method because nearest neighbors-based techniques require a similarity measure.



(a) Conventional implementation: only pairs of aligned samples.



(b) Weighting method exemplified for samples p_{10} and q_6: all possible pairs of samples, with each pair weighted according to time misalignment. Larger weights are represented by darker tones.

Figure 2. Pairs of differences (represented by arrows) used in assessing the distance between segments $\mathbf{p}$ (black line and markers) and $\mathbf{q}$ (grey line and markers).

The conventional implementation of distance and correlation measures is illustrated in Figure 2a. For the case of the distance measure, the arrows in the figure indicate that only differences between aligned samples are considered. Instead, the weighting method calculate differences between all possible pairs of samples, and weights each difference according to the time misalignment between the pair (Rehfeld et al., 2011). This idea is illustrated in Figure 2b for sample p_{10} and sample q_6. Larger weights are represented in the figure by darker tones on the arrows. The weighting function is such that the more aligned samples are, the more their difference counts towards the distance metric.

The weighted version of the Euclidean distance metric is defined as

$$d(\mathbf{p},\mathbf{q},w) = \sqrt{\sum_{i=1}^{n_p}\sum_{j=1}^{n_q} w_{i,j}(p_i - q_j)^2} \tag{2}$$

where $w_{i,j}$ is the weight atributed to the difference between sample p_i of time series $\mathbf{p}$ and sample q_j of time series $\mathbf{q}$. Examples of weight functions w found in the literature are sinc and Gaussian functions (Rehfeld et al., 2011). In particular, the Gaussian function (equation (3)) is a positive function which decays smoothly to zero, and is symmetric with relation to the time misalignment $(t_i - t_j)$ between samples p_i and q_j.

$$w_{i,j} = w(t_i, t_j) = \frac{1}{\sqrt{2\pi}L} \exp\left(\frac{-(t_i - t_j)^2}{2L^2}\right) \quad (3)$$

Since a distance metric should be non-negative and symmetric, the Gaussian function is a relevant alternative for a weighting function. The Gaussian weighting function has a width parameter L which determines the rate of decay of the weight values $w_{i,j}$ with the time misalignment between the two samples.

Other methods to analyse irregularly sampled time series include: (i) reconstruction methods, (ii) spectral transforms, and (iii) ARMA model fitting (Rehfeld et al., 2011).

Reconstruction methods resample the time series into a regular time grid and then apply existing methods developed for regularly sampled time series. Common techniques of resampling include linear and spline interpolation, regression, and approximation by the value of the sample closest in time (Lall & Sharma, 1996).

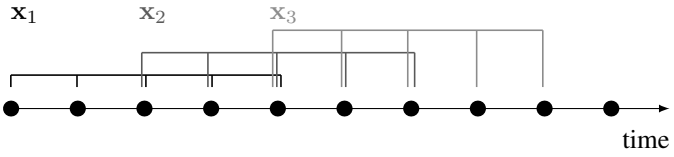
A common spectral transform for irregularly sampled time series is the Lomb-Scargle Fourier transform (Scargle, 1989). It determines the spectrum of a measurement from a least squares fit of sine curves to the time series of the measurement. It is suitable for measurements with periodic components and no outliers (Stoica et al., 2009). The wavelet transform can also be computed for irregularly sampled time series if implemented through the lifting scheme (Sweldens, 1998).

Fitting autoregressive-moving-average (ARMA) models to a time series involves determining the coefficients of the ARMA model. To determine the coefficients from irregularly sampled time series, research focuses on adapting estimation algorithms such as maximum-likelihood estimation (Isaksson, 1993) and the Burg algorithm (Bos et al., 2002).

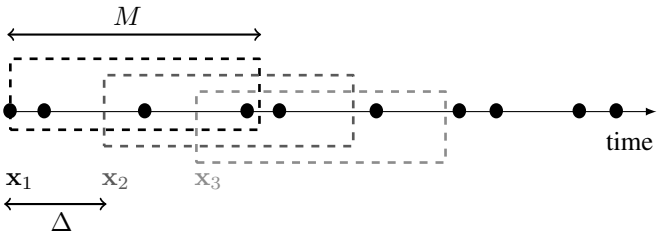
2.2. Time-based construction of embedded vectors

Embedded vectors were originally defined for regularly sampled time series (Kantz & Schreiber, 2003). They refer to segments from a time series with a fixed number m of samples, with each embedded vector lagging the previous by δ samples (Kantz & Schreiber, 2003). Figure 3a illustrates the selection of three embedded vectors from a symbolical time series represented by dots, with $m = 5$ and $\delta = 2$. Embedded vectors are commonly used in the analysis of nonlinear time series.

Cecílio et al. (2015) proposed an alternative approach to the construction of embedded vectors motivated by the integrated analysis of measurements with fast and slow sampling rates. That paper imposed the same time span M for all embedded vectors and a lag of a constant number Δ of time units for the different measurements. This way, the embedded vectors of the different measurements would be synchronized.



(a) Time series X is regularly sampled. Embedded vectors have $m = 5$ and $\delta = 2$.



(b) Time series X is irregularly sampled. Each embedded vector spans M time units and lags the previous by Δ time units.

Figure 3. Representations of a time series X and the construction of first three embedded vectors.

This paper uses the same idea because if a fixed number m of samples were imposed to an irregularly sampled time series, then the embedded vectors would not span the same duration of time. If the same step δ were imposed, then embedded vectors of different measurements would not be aligned. The difference to the problem in Cecílio et al. (2015) is that with irregular sampling rates, the use of a constant time span M implies that the embedded vectors of a single measurement may have different numbers of samples. The weighted Euclidean distance metric discussed previously is able to compute distances between embedded vectors with these sampling characteristics.

3. METHODS

This section explains the techniques to construct embedded vectors and to compute similarity in the case of a measurement with irregular sampling rate. These techniques extend the applicability of time series analysis methods based on nearest neighbors of embedded vectors.

3.1. Embedded vectors

Consider a time series X of sample values $x(t_i)$ (equation (4a)) which are ordered according to the time sequence T of strictly increasing sampling instants t_i (equation (4b)).

$$X = \{x(t_1), x(t_2), \cdots, x(t_n)\} : t_1 < t_2 < \cdots < t_n \quad (4a)$$

$$T = \{t_1, t_2, \cdots, t_n\} : t_1 < t_2 < \cdots < t_n \quad (4b)$$

An embedded vector $\mathbf{x}_r$ is defined as a segment of the time

series X which spans M time units. The number of samples is variable and is here denoted as m_r. Furthermore, embedded vector $\mathbf{x}_r$ lags the previous $\mathbf{x}_{r-1}$ by a constant number Δ of time units. The construction of embedded vectors from X is represented in Figure 3b.

Additionally, for each embedded vector $\mathbf{x}_r$, a time vector $\mathbf{t}_r$ should be created to arrange the time instants of each sample in $\mathbf{x}_r$, that is, $\mathbf{t}_r = \{t_{r,1}, t_{r,2}, \cdots, t_{r,m_r}\}$.

It should be noted that the embedded vectors of a measurement cannot be arranged in an embedding matrix, as conventionally done with regularly sampled measurements (Thornhill, 2005; Cecílio et al., 2014). This is due to the different number of samples in each embedded vector, as illustrated in Figure 3b.

3.2. Similarity

Each pair of embedded vectors $\mathbf{x}_r$ and $\mathbf{x}_s$ is then compared using the weighted Euclidean distance metric

$$d(\mathbf{x}_r, \mathbf{x}_s, w)^s = \frac{\sqrt{\sum_{i=1}^{m_r}\sum_{j=1}^{m_s} w_{i,j}(x_{r,i} - x_{s,j})^2}}{\sqrt{\sum_{i=1}^{m_r}\sum_{j=1}^{m_s} w_{i,j}}} \quad (5)$$

where i and j represent the indices of the samples in $\mathbf{x}_r$ and $\mathbf{x}_s$, respectively. This equation is a scaled version of the weighted Euclidean distance metric presented in equation (2). The aim of scaling is to have a metric $d(\mathbf{x}_r, \mathbf{x}_s, w)^s$ which is independent of the number of samples in $\mathbf{x}_r$ and $\mathbf{x}_s$.

The weighting function $w_{i,j} = w(t_{r,i} - t_{s,j})$ is defined as in equation (3), and depends on the time instants of the samples of $\mathbf{x}_r$ and $\mathbf{x}_s$. The width parameter L may be optimized or used as suggested in (Rehfeld et al., 2011), that is,

$$L = \frac{\bar{\Delta}t}{4} \quad (6)$$

where $\bar{\Delta}t$ is the mean value of the sampling intervals in measurement X.

4. APPLICATION TO UNIVARIATE DETECTION OF TRANSIENT DISTURBANCES

This section demonstrates the proposed formulation of embedded vectors and similarity measure within a method for extracting the moment of occurrence as well as intensity of transient disturbances (Cecílio et al., 2014). In electrical circuits, transient disturbances include voltage spikes, which can be an indication of unbalanced power grid as well as a cause of degradation of sensitive electronics (Bevrani, 2009). In rotating tools, transient disturbances can indicate abnormal shock and vibration levels.

4.1. Univariate detection of transient disturbances

In Cecílio et al. (2014), transient disturbances were formally defined as infrequent and short-lasting deviations of a measurement from its underlying trend. The extraction of the moment of occurrence and intensity of these features was formulated as an anomaly detection problem, and solved using nearest neighbors of embedded vectors.

The implementation of the method can be summarized in the following steps.

1. Embedded vectors with a fixed number m of samples and δ samples of lag are generated from a time series X.
2. Each embedded vector is then compared to every other embedded vector, using the Euclidean distance metric.
3. An anomaly index ai is then attributed to each embedded vector $\mathbf{x}_r$ as the k^{th} smallest distance between $\mathbf{x}_r$ and every other embedded vector. This is denoted by d_k , the distance to its k^{th} nearest neighbor.
4. An anomaly index vector $\mathbf{ai}$ is formed by the sequence of anomaly indices ai of each embedded vector.
5. A threshold based on the statistics of $\mathbf{ai}$ distinguishes embedded vectors which capture the transients from embedded vectors which capture periods of normal operation.

In the following, the two initial steps of the method will be replaced by the alternative formulation of embedded vectors and similarity proposed in section 3. With this formulation the method of univariate detection of transient disturbances can now be applied to irregularly sampled time series.

4.2. Case study

The proposed method uses a measurement from Cecílio et al. (2014) in order to have a benchmark for the results. The measurement represents the shaft speed of a compressor during 20 seconds, and was obtain from a gas compressor rig located at ABB Corporate Research Center, Kraków, Poland. The shaft speed was measured at a regular rate of 1 kHz, therefore its measurement had to be manipulated in order to have an irregularly sampled time series for illustration of the proposed method. This was done by randomly eliminating samples from the original measurement, which also resulted in a decrease in the total number of samples from 20,000 samples to approximately 400. The time instants of the retained samples were stored. Figure 4 shows the speed measurement after this manipulation. Figure 5 shows a close-up to highlight the irregular spacing between samples.

Two step changes, around 5 and 11 s, were imposed in the drive of the compressor by changing its speed set-point, resulting in the two transients seen in the figure. The objective of the proposed method is to detect those transients.

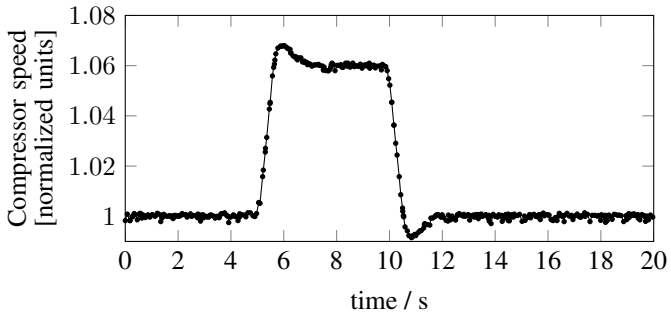


Figure 4. Compressor speed measurement from (Cecílio et al., 2014). The measurement was manipulated in order to have an irregularly sampled time series. The values are normalized by the initial value.

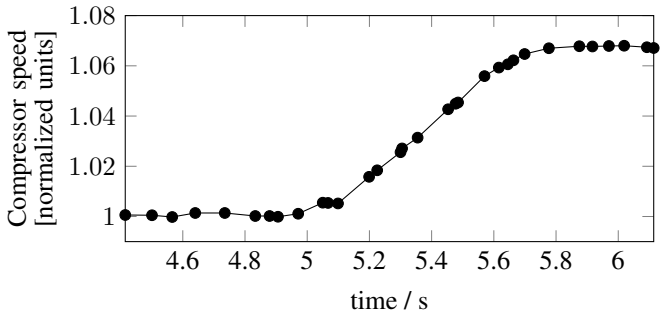


Figure 5. Close-up on the compressor speed measurement to highlight the irregular spacing between samples.

4.3. Results

Figure 6a shows the anomaly index vector $\mathbf{ai}$ computed from the measurement in Figure 4. As in Cecílio et al. (2014), $\mathbf{ai}$ was normalized by its median so that $ai = 1$ now approximates the average anomaly index of non-anomalous embedded vectors.

The positive detection of the two transients is indicated by the fact that the embedded vectors which correspond to transient disturbances have anomaly indices above the detection threshold, which is represented by the dashed line in Figure 6a. The figure shows that the construction of embedded vectors and similarity measure suggested in section 3 are able to cope with the sampling irregularity and achieve the desired detection.

Figure 6b shows the result obtained by Cecílio et al. (2014) with the original regularly sampled measurement. The figure clearly shows the similarity between the two results. This demonstrates the potential of the proposed formulation for the analysis of irregularly sampled time series with nearest neighbor-based methods.

4.4. Comment on the use of the method for real-time monitoring

The techniques proposed in this paper to construct embedded vectors and to compute similarity are applicable in both on-line and off-line analysis methods. In section 4, the techniques were demonstrated as part of an off-line analysis because the transients detection method proposed in Cecílio et al. (2014) was originally implemented in that way. However, the concept of transients detection with nearest neighbors is amenable to on-line implementation, and this section discusses possible approaches.

The crucial point for on-line implementation is the computational cost of comparing every embedded vector to all other embedded vectors.

One way to reduce this cost is the following. When new samples arrives and a new embedded vector is formed, that embedded vector is compared against a finite number N_H of past embedded vectors. This amounts to N_H computations of distance between two vectors. The anomaly index ai of that embedded vector comes from one search operation amongst those N_H distances.

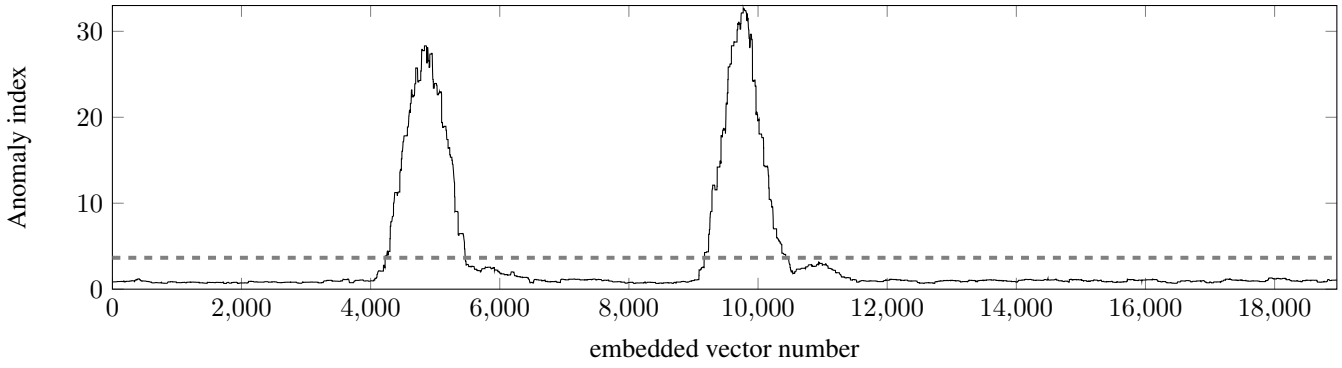
The efficiency of the distance computations can also be improved. The anomaly index ai only uses one piece of information out of the N_H distances calculated for each embedded vector. However, these N_H distances can also be used to identify tight clusters of embedded vectors. As a result, every time a new embedded vector is formed it needs only be compared to the centroid of each cluster instead of all the embedded vectors that form that cluster.

These modifications in the implementation should enable the on-line implementation of the transients detection, which is better suited for PHM applications.

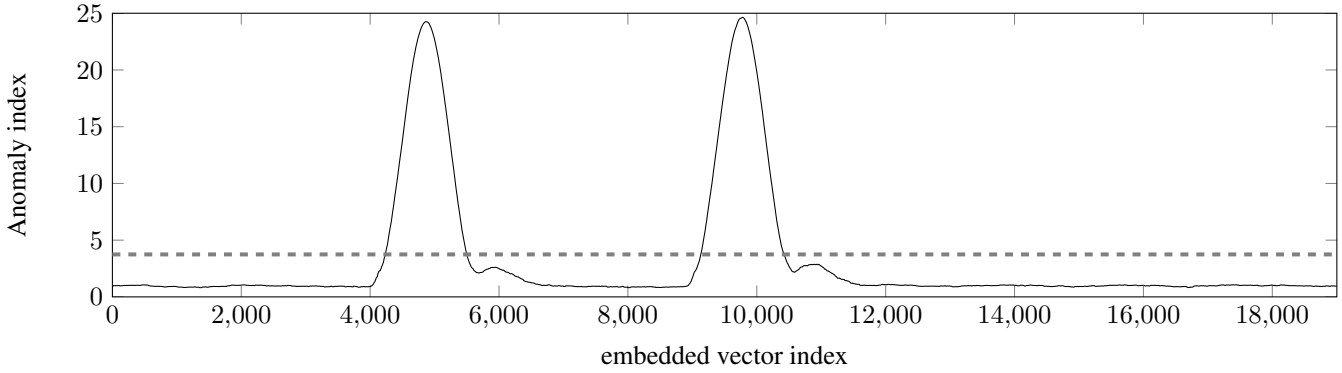
5. CONCLUSIONS

This paper presented an adaptation to methods based on nearest neighbors to enable their application to measurements with irregular sampling rates. The first two steps of these methods normally involve the construction of embedded vectors and a similarity assessment. With irregular sampling rates the conventional construction of embedded vectors and similarity measure cannot be applied. The proposed techniques comprise a time-based formulation for embedded vectors, and a weighted distance metric to assess the similarity between the embedded vectors. These techniques can substitute the first two steps of conventional nearest neighbors methods.

The new techniques were demonstrated within a method of detection of transient or spiky disturbances, which had been developed for regularly sampled measurements. The case study showed that the new formulation achieves results in an irregularly sampled time series on a par with the results ob-



(a) Obtained with the irregularly sampled measurement and the method proposed.



(b) Obtained with the regularly sampled measurement and the original method.

Figure 6. Normalized anomaly index vector. The dashed line indicates the detection threshold.

tained with the original regularly sampled measurement. This supports the research potential of the idea proposed in this paper.

Open questions about the proposed idea include:

- studying if, and under which conditions, the weighted Euclidean metric converges to conventional Euclidean metric,
- determining the statistical behaviour of the anomaly index vectors in order to attribute a confidence level to the selected threshold, in the case of the detection methods,
- optimizing the width parameter L, and re-evaluating the parameter optimisation done for methods with regularly sampled measurements, and
- analysing the sensitivity of the methods to the distribution of samples in the measurement.

The paper also discussed possible modifications to enable the on-line implementation of the techniques.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from the Portuguese Foundation for Science and Technology (FCT) under Fellowship SFRH/BD/61384/2009 and the Marie Curie FP7-IAPP project "REAL-SMART - Using real-time measurements for monitoring and management of power transmission dynamics for the Smart Grid", Contract No: PIAP-GA-2009-251304.

REFERENCES

Babji, S., & Tangirala, A. K. (2010). Source separation in systems with correlated sources using NMF. *Digital Signal Processing*, *20*(2), 417–432.

Bauer, M., & Thornhill, N. F. (2008). A practical method for identifying the propagation path of plant-wide disturbances. *Journal of Process Control*, *18*(7–8), 707–719.

Bevrani, H. (2009). *Robust power system frequency control*. Springer.

Bos, R., de Waele, S., & Broersen, P. M. T. (2002). Autoregressive spectral estimation by application of the Burg algorithm to irregularly sampled data. *IEEE Transactions on Instrumentation and Measurement*, *51*(6), 1289–1294.

Cecílio, I. M., Ottewill, J. R., Fretheim, H., & Thornhill, N. F. (2015). Multivariate detection of transient disturbances for

uni- and multirate systems. *IEEE Transactions on Control System Technology*, *23*(4), 1477–1493.

Cecílio, I. M., Ottewill, J. R., Pretlove, J., & Thornhill, N. F. (2014). Nearest neighbors method for detecting transient disturbances in process and electromechanical systems. *Journal of Process Control*, *24*(9), 1382–1393.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58.

Choudhury, M. A. A. S., Shah, S. L., & Thornhill, N. F. (2004). Diagnosis of poor control-loop performance using higher-order statistics. *Automatica*, *40*(10), 1719–1728.

Isaksson, A. J. (1993). Identification of ARX – models subject to missing data. *IEEE Transactions on Automatic Control*, *38*(5), 813–819.

Kantz, H., & Schreiber, T. (2003). *Nonlinear time series analysis*. Cambridge University Press.

Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, *32*(3), 679–693.

Rehfeld, K., Marwan, N., Heitzig, J., & Kurths, J. (2011). Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics*, *18*(3), 389–404.

Russell, E. L., Chiang, L. H., & Braatz, R. D. (2000). *Data-driven methods for fault detection and diagnosis in chemical processes* (1^{st} ed.). Springer.

Scargle, J. D. (1989). Studies in astronomical time series analysis. III. Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data. *Astrophysical Journal*, *343*, 874–887.

Stockmann, M., Haber, R., & Schmitz, U. (2012). Source identification of plant-wide faults based on k nearest neighbor time delay estimation. *Journal of Process Control*, *22*(3), 583–598.

Stoica, P., Li, J., & He, H. (2009). Spectral analysis of nonuniformly sampled data: a new approach versus the periodogram. *IEEE Transactions on Signal Processing*, *57*(3), 843–858.

Sweldens, W. (1998). The lifting scheme: A construction of second generation wavelets. *SIAM Journal on Mathematical Analysis*, *29*(2), 511–546.

Tangirala, A. K., Kanodia, J., & Shah, S. L. (2007). Non-Negative matrix factorization for detection and diagnosis of plantwide oscillations. *Industrial & Engineering Chemistry Research*, *46*(3), 801–817.

Thornhill, N. F. (2005). Finding the source of nonlinearity in a process with plant-wide oscillation. *IEEE Transactions on Control Systems Technology*, *13*(3), 434–443.

Thornhill, N. F., Shah, S. L., Huang, B., & Vishnubhotla, A. (2002). Spectral principal component analysis of dynamic process data. *Control Engineering Practice*, *10*(8), 833–846.

Zang, X., & Howell, J. (2007). Isolating the source of whole-plant oscillations through bi-amplitude ratio analysis. *Control Engineering Practice*, *15*(1), 69–76.

Zumbach, G., & Müller, U. (2001). Operators on inhomogeneous time series. *International Journal of Theoretical and Applied Finance*, *4*(01), 147–177.