Optimizing Pervasive Sensor Data Acquisition utilizing Missing Values Substitution

M. Michalopoulos
Faculty of Informatics
Hellenic Open University,
Greece
mmgrev@yahoo.com

C. Anagnostopoulos
Dept. of Informatics and Telecommunications,
University of Athens,
Ilissia, Athens, Greece
bleu@di.uoa.gr

Charalampos Doukas
University of the Aegean
Dep. of Information & Communication Systems Engineering
Samos, Greece
doukas@aegean.gr

Ilias Maglogiannis
University of Central Greece
Dep. of Computer Science & Bioinformatics
Lamia, Greece
imaglo@ucg.gr

S. Hadjiefthymiades
Dept. of Informatics and Telecommunications
University of Athens
Ilissia, Athens, Greece
shadj@di.uoa.gr

ABSTRACT
Acquisition of pervasive sensor data can be often unsuccessful due to power outage at nodes, time synchronization issues, interference, network transmission failures or sensor hardware issues. Such failures can lead to inadequate data delivery to the monitoring applications resulting in erroneous conclusions. This paper presents a missing values substitution framework that addresses the aforementioned issue. The presented framework has been evaluated within a pervasive sensor monitoring environment that collects and transmits patient health related data and results are presented.

Categories and Subject Descriptors
H.4.3 [Communications Applications], J3.3 [Medical information systems]

General Terms
Algorithms, Measurement, Performance, Reliability, Experimentation.

Keywords
Pervasive Sensors, Healthcare Data Transmission, Missing Values Substitution.

1. INTRODUCTION
Pervasive environments offer improved living conditions and levels of independence for patients and the elderly population who require support with both physical and cognitive functions. Within these environments sensing technologies provide a key facility to monitor the behavior of the person and their interactions. Wireless technologies enable the real time transmission of data about a patient's condition to caregivers. Numerous portable devices are available that can detect certain medical conditions—pulse rate, blood pressure, breath alcohol level, and so on—from a user's touch. Many such capabilities already have been integrated into a handheld wireless device that also contains the user's medical history. It may even be possible to detect certain contextual information, such as the user's level of anxiety, based on keystroke patterns. After analyzing data input, the device could transmit an alert message to a healthcare provider, the nearest hospital, or an emergency system if appropriate.

In order for such systems to be efficient and effective, the data obtained from sensors within the monitoring environment have to be totally reliable and robust. However, this cannot be achieved in real life monitoring due to a number of reasons, such as time synchronization issues, interference, network transmission failures in a wireless sensors network.

Within this context, this paper presents a Fuzzy Identification System (FIS) and a Recursive Probabilistic Principal Component Analysis (RPPCA), for dealing with missing values derived from data streams of wireless bio-sensor networks. The rest of the paper is structured as follows: Section 2 discusses related work in the domain of missing values substitution for sensor networks. Section 3 presents an overview of pervasive sensor data acquisition methods and techniques in healthcare. Section 4 describes the proposed methods and Section 5 presents evaluation results using an on-body pervasive sensor platform. Finally, Section 6 concludes the paper and report future trends.

2. RELATED WORK
The problem of missing values appears in applications of many areas. This problem is one of the failures that emerge more frequently in wireless sensor networks. Many approaches have been proposed for the substitution of missing values. The most popular approach is the use of the mean substitution. We can use many other simple or complex univariate statistical methods like the mean substitution or autoregressive moving average (ARMA) ([16]). We

PETRA'10, June 23 - 25, 2010, Samos, Greece.

can also use multivariate statistical methods like principal component analysis (PCA) or partial least squares (PLS) regression for correlated data ([16]).

Today's sensor networks can deploy easily, are low-budget and produce redundancy of correlated data. We can exploit the last attribute to estimate missing values using multivariate techniques. Wasito et al. ([18]) discussed and compared different approaches for missing values imputation. Dunia et al. ([19]) countered missing values like other faults found in process data, e.g. outliers or complete failures using PCA. PCA can reduce the dimensions of a sample by converting the parameters of measurement to a new system of axes ([17]). The new parameters are linear combination of the corresponding initial ones. Nelson et. al. ([20]) proposed three methods based on PCA and PLS. Arteaga et al. ([21]) stated a few more approaches and proved that some of these are equivalent to methods in [20]. The authors of [19], [20] and [21] intended to calculate the scores in the presence of missing values and use them in monitoring systems. If the scores are calculated, the missing values can be easily extracted.

A solution of missing values problem is to use the maximum likelihood estimation (MLE) to provide estimates of missing values. We can find MLE of the parameter vector using the expectation – maximization (EM) algorithm ([22]) that is a simple but computationally complex iterative procedure.

3. PERVASIVE SENSOR DATA ACQUISITION IN HEALTHCARE

This section presents an overview of techniques and methods for acquiring medical data through pervasive sensors. The most popular biosignals utilized in pervasive health applications ([1]-[10]) are summarized in the table below.

Table 1. Broadly used biosignals with corresponding metric ranges, number of sensors required and information rate.

Biomedical Measurements	Voltage range (V)	Number of sensors	Information rate (b/s)
ECG	0.5-4 m	5-9	15000
Heart sound	Extremely small	2-4	120000
Heart rate	0.5-4 m	2	600
EEG	2-200 μ	20	4200
EMG	0.1-5 m	2+	600000
Respiratory rate	Small	1	800
Temperature of body	0-100 m	1+	80

In addition to the aforementioned biosignals, patient physiological data (e,g., body movement information based on accelerometer values), and context-aware data (e.g., location, environment and age group information) have also been used by pervasive health applications ([9], [10]). In the context of pervasive healthcare applications, the acquisition of biomedical signals is performed through special devices (i.e. sensors) attached on the patients body (see Figure 1) or special wearable devices (see Figure 2).

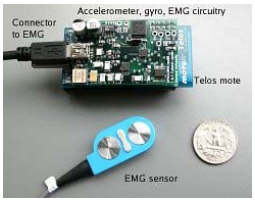
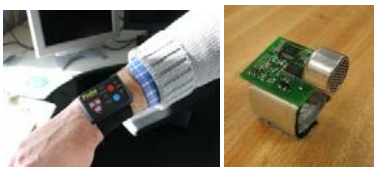


Figure 1. Accelerometer, gyroscope, and electromyogram (EMG) sensor for stroke patient monitoring [11].



(a) (b)

Figure 2. Wearable medical sensor devices: (a) A 3-axis accelerometer on a wrist device enabling the acquisition of patient movement data [11], (b) A ring sensor for monitoring of blood oxygen saturation [12].

Regarding communication, there are two main enabling technologies according to their topology: on-body (wearable) and off-body networks. Recent technological advances have made possible a new generation of small, powerful, mobile computing devices. A wearable computer must be small and light enough to fit inside clothing. Occasionally, it is attached to a belt or other accessory, or is worn directly like a watch or glasses. An important factor in wearable computing systems is how the various independent devices interconnect and share data. An off-body network connects to other systems that the user does not wear or carry and it is based on a Wireless Local Area Network (WLAN) infrastructure, while an on-body or Wireless Personal Area Network (WPAN) connects the devices themselves; the computers, peripherals, sensors, and other subsystems and runs at ad hoc mode. WPANs are defined within the IEEE 802.15 standard. The most relevant protocols for pervasive e-health systems are Bluetooth and ZigBee. The latter has been developed as a low data rate solution with multi-month to multiyear battery life and very low complexity. It is intended to operate in an unlicensed international frequency band. The maximum data rates for each band are 250, 40, and 20 kbps, respectively.. Finally, 3G mobile connectivity offers the freedom to leave the home and access high data rate services, even video using readily available and low-cost devices.

Mobility is another major issue for pervasive e-health applications because of the nature of users and applications and the easiness of the connectivity to other available wireless networks. Both off-body and personal area networks must not have line-of-sight (LoS) requirements. The various communication modalities can be used in different ways to construct an actual communication network. Figure 3 illustrates a general architecture scheme based on the aforementioned, for enabling delivery of patient medical data to caregivers and physicians through pervasive sensors.

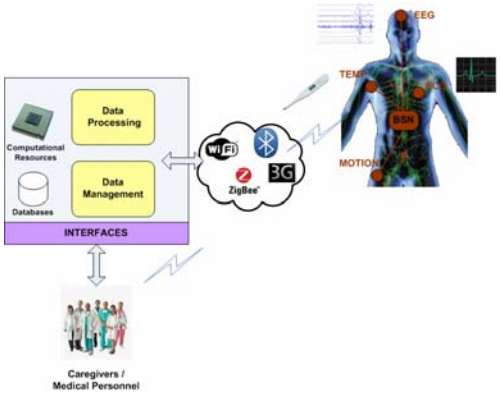


Figure 3. General architecture scheme of medical data acquisition in a pervasive environment.

During the transmission of medical and other patient-related data several failures might occur that can affect the assessment and diagnosis:

- Power failure at nodes: Usually body sensor networks are characterized by low power consumption devices. However, the sensor nodes continuously operate collecting medical data and transmitting them wirelessly to the monitoring units. The latter process requires enough power resources otherwise transmission can become weak and introduce pauses or failures in the delivery of the data streams.
- Hardware failures at biosignal monitoring interfaces: The sensor nodes contain interfaces that are connected to a variety of monitoring devices (e.g., ECG or EEG circuits). There can be often cases where these interfaces can malfunction or pause acquiring data (e.g., due to lead accidental removal by the patient), or noise can be interfered to data streams (e.g., noise introduced by patient movement).
- Additional transmission failures can happen due to:
    - Range issues: The most common communication protocols for body sensor networks (i.e. Bluetooth and ZigBee) have an average range of 10-50 meters that can be affected by several parameters, like intermediate objects, power, etc. It is very common when patient moves around in the monitoring environment that the sensor nodes can get temporarily out of range of the receivers and cause missing values in the data streams.
    - Interference issues: Similar data failures can be caused due to interferences in the wireless medium by several devices that exist in the monitoring environments and emit electromagnetic fields (e.g., mobile phones, TV screens, etc.).
    - Synchronization issues: Sensor nodes vastly acquire and transmit different kinds of patient data with different sampling frequencies. The latter can cause synchronization issues to the receiving applications leading to missing or false interpreted data values.

All the aforementioned situations can result in missing or false values in the received data streams. The following sections presents the proposed methods for addressing such issues.

4. MISSING VALUES SUBSTITUTION

Missing values in a data stream occur when no value is reported (e.g., a sensor reading) for one or more specific variables in a current observation. Missing values can lead to erroneous conclusions about data and, in turn, erroneous inference on the whole situation / context of an entity. In addition, missing values may prevent proper classification of the situation of an entity, and poor substitution schemes for missing values may cause classification errors. If all the values substituted are determined by the most likely value, then the individual values are less likely to help in situational inference. Similarly, substitution of missing values may introduce inaccuracies and inconsistencies. Missing data values can negatively impact classification results, and errors or data skews can proliferate across subsequent runs and cause a larger, cumulative error, etc. Moreover, most analysis methods cannot be performed if there are missing values in a data stream. Generally speaking, we can consider the two policies for resolving missing values:

- Eliminate observations with a high number of missing values, since estimating high numbers of missing values may introduce bias to further analysis.
- Replace the observations with missing values adopting missing value substitution techniques like, estimating missing values by a measure of central tendency, by nearest neighbors, by replacing missing values with an arbitrary value.

Both policies can be adopted regarding the application. For instance, whether a critical application has to detect and infer the situation of an elderly person in a home-care environment then each observation derived from a sensors stream is of high importance (e.g., ECG). On the other hand, whether an application attempts to control a heat system in a medical room deriving information from a wireless sensors network (including temperature and humidity sensors) then the former policy is preferable.

In this work, we investigate two methods, a Fuzzy Identification System (FIS) and a Recursive Probabilistic Principal Component Analysis (RPPCA), for dealing with missing values derived from data streams of wireless bio-sensor networks. Specifically, we target on critical applications, which regularly require sensing information in order to detect and infer the situations of elderly persons in a home-care environment. Actually, such applications are in need of full of contextual information on the current situation thus any adopted model for missing value substitution has to assume a low computational cost.

4.1 On the use of Fuzzy Identification System

Let $\mathbf{X} = [X_1, \ldots, X_n]^T$ be a n-dimensional context vector of n contextual parameters. We plan to estimate the missing value of the X_j contextual parameter for some j=1, …, n. If we have a series of m observations (measurements) of $\mathbf{X}$ written as $\mathbf{X}^k$, k=1, …, m and the (m + 1)-th value x_j^{m+1} of X_j is missing then we estimate x_j^{m+1} based on the m observations $[x_1^k, \ldots, x_n^k]$, k=1, …, m and the m+1 observation $[x_1^{m+1}, \ldots, x_{j-1}^{m+1}, x_{j+1}^{m+1}, \ldots, x_n^{m+1}]$. We calculate the *sample correlation coefficient* matrix r_{ij} for each pair of contextual parameters (X_i, X_j), that is,

$$r_{ij} = \frac{\sum_{l=1}^{m} \left(x_i^l - \bar{x}_i\right)\left(x_j^l - \bar{x}_j\right)}{(m-1)\sigma_i \sigma_j}$$

where $\bar{x}_i, \bar{x}_j$ are the sample means, and σ_i, σ_j are the sample standard deviations of X_i, and X_j, respectively. The absolute value of the sample correlation must be less than or equal to 1. We select those X_i that are most correlated with the missing X_j for estimating x_j^{m+1}. Hence, if $\mathbf{X}_D$ is the context vector with those contextual parameters that are most correlated with the missing parameter w.r.t. r then

$$\mathbf{X}_D = \mathbf{X}^T \, diag(d_1, \ldots, d_n)$$

where $diag(d_1, \ldots, d_n)$ is a diagonal matrix with $d_i = 1$ if $|r_{ij}| > \varepsilon$ for some $\varepsilon > 0$; otherwise $d_i = 0$. We call the elements of $\mathbf{X}_D$ as *regressors* (or explanatory context) that attribute to some linear combination on X_j. This means that, the $n_D \leq n$ -1 contextual attributes of $\mathbf{X}_D$ can be used for estimating the missing value of X_j after a linear regression of the n_D contextual attributes on X_j.

Moreover, we assign for each observation an exponential weighted factor for putting less emphasis on old context vectors and importance on the recent observations. Let W be a diagonal matrix $m \times n_D$ with its diagonal elements $w_k = \lambda^{m-k}$ for k=1, …, m, m, $0 < \lambda < 1$ and its off-diagonal elements equal to zero. Hence, the most important observation is the m-th with $w_m = 1$ and for the k =1, …, m-1 observations it holds that $w_1 < \ldots < w_{m-1}$. The weight of importance over the m observations is $(1 - \lambda)^{-1}$, e.g., for $\lambda = 0.8$, the fifth most recent observations are the most important for determining the missing value of X_j.

We adopt the Weighted Least Squares (WLS) method as a method for linear regression of $\mathbf{X}_D$ on $Y = X_j$ that determines the parameter vector θ of unknown quantities in a statistical model by minimizing the sum of the squared difference between the y^k and observation $[x_1^{m+1}, \ldots, x_{j-1}^{m+1}, x_{j+1}^{m+1}, \ldots, x_n^{m+1}]$, $k = 1, \ldots, m$ values. The estimate of θ that results to a best fit to the set of observations, that is, $Y = \mathbf{\Phi} \cdot \theta + E$, is then

$$\theta' = (\mathbf{\Phi}^T W \mathbf{\Phi})^{-1} \mathbf{\Phi}^T W Y$$

E is the random errors e_k, k=1, …, m that are with zero expected value, uncorrelated, have the same variance and are independent of $\mathbf{X}_D$. The $m \times n_D$ matrix $\mathbf{\Phi}$ can be the matrix of the m observations, i.e., $\mathbf{\Phi} = [(\mathbf{X}_D^k)^T]$, k=1, …, m. However, $\mathbf{\Phi}$ can be a mapping of the context observations to specific regression vectors resulted from the identification of a fuzzy system. Specifically, we tune a fuzzy system f with input the $\mathbf{X}_D$ context vector and output the X_j contextual parameter for estimating the missing value of x_j at the (m+1)-th observation.

We consider f as a nonlinear mapping between the n_D inputs and the y output [25]. The inputs are crisp, i.e., they are real numbers (not fuzzy sets). The *fuzzification* process converts the crisp inputs into fuzzy sets, the *inference mechanism* uses the fuzzy rules in the rule-base to produce fuzzy conclusions, and the *defuzzification* process converts these fuzzy conclusions into the crisp outputs. We tune f by estimating the *best* parameters for the regression process. We begin by precisely defining the *function approximation problem*, in which we seek to construct a function f to approximate another function g that is inherently represented by a finite number of input-output associations, i.e., regressors-missing parameter associations. We construct a nonlinear estimator incorporating fuzzy variables for the contextual parameters in $\mathbf{X}_D$. Given the *real* function on the observations g: $\mathbf{X}_D{}^r \rightarrow Y^r$, where $\mathbf{X}_D{}^r \subset X_1 \times \ldots \times X_{nD}$, we construct a fuzzy system f: $\mathbf{X}_D \rightarrow Y$, $\mathbf{X}_D \subset \mathbf{X}_D{}^r$ and $Y \subset Y^r$ are domain and range of interest, by choosing a parameter vector θ, which includes membership function centers and widths of the output, so that,

$$g(x) = f(x \mid \theta) + e(x)$$

for all $x = [x_1, \ldots, x_{nD}]^T \in \mathbf{X}_D$ where the approximation error $e(x)$ is as small as possible. All that is available to choose the parameters θ of the fuzzy system $f(x \mid \theta)$ is some part of the function g in the form of a finite set of input-output data pairs (x^i, y^i), $x^i \in \mathbf{X}_D$, $y^i \in Y$ and $y^i = g(x^i)$. If $x^i = [x_1{}^i, \ldots, x_{nD}{}^i]^T$ represents the input vector for the i-th data pair, then the training data set of m data pairs is denoted by $G(m) = \{(x^1, y^1), \ldots, (x^m, y^m)\} \subset \mathbf{X}_D \times Y$. Hence, the Fuzzy Identification System that is constructed is given by

$$f(x) = \frac{\sum_{i=1}^{m} b_i \mu_i(x)}{\sum_{i=1}^{m} \mu_i(x)} \tag{1}$$

where $\mu_i(x)$ is the certainty of the premise of the i-th rule (i.e., the i-th data pair) specified by the membership functions on the input domain, that is,

$$\mu_i(x) = \prod_{j=1}^{n_D} \exp\left(-\frac{1}{2}\left(\frac{x_j - c_j^i}{\sigma_j^i} \right)^2 \right)$$

f uses singleton defuzzification, Gaussian membership functions, product for the premise and implication, and center-average defuzzification.

We use the WLS method to tune the fuzzy system f and estimate the centers of the output membership functions (the resulted parameter vector θ), b_i, i =1, …, m, that is the $\theta = [b_1, \ldots, b_m]$ parameter. Note that, if $\xi_i(x) = \dfrac{\mu_i(x)}{\sum_{j=1}^{m} \mu_j(x)}$ then $f(x \mid \theta) = b_1\xi_1(x) + \ldots b_m\xi_m(x)$. Hence, if we define $\xi(x) = [\xi_1(x), \ldots, \xi_m(x)]^T$, then $y = f(x \mid \theta) = \theta^T \xi(x)$. Evidently, if the μ_i are given then $\xi(x)$ is given so that it is in exactly the right form for use by the WLS method as long as $\xi(x)$ is viewed as a regression vector. The WLS algorithm produces an estimate $\theta' = (\mathbf{\Phi}^T W \mathbf{\Phi})^{-1} \mathbf{\Phi}^T W Y$ of the best centers for the output membership function centers b_i with

$$\mathbf{\Phi} = \mathbf{\Phi}(m) = [(\xi(x^1))^T, \ldots, (\xi(x^m))^T]^T$$

Hence, the fuzzy sets for the regression process are parameterized as *linear in the parameters* via the mapping of x^i to $\xi(x^i)$ and tuned to achieve perfect estimation of θ minimizing the error $E = Y - \mathbf{\Phi} \cdot \theta$. The estimation of the (m+1)-th value of the missing parameter X_j, i.e., $y = y^{m+1}$, is then derived from $y = f(x)$ in Equation (1) putting $x = x^{m+1} = [x_1{}^{m+1}, \ldots, x_{nD}{}^{m+1}]^T$.

It should be noted that, to tune f with may outputs, i.e., more than one contextual parameters X_j are missing, then we simply repeat the algorithm described below for each output (missing parameter) with the following steps:

1. Let $J = \{j_1, \ldots, j_l\}$, $l < n$ be the set of indices that correspond to the missing contextual parameters $\{X_{j1}, \ldots, X_{jl}\}$.

2. We select the missing contextual parameter X_{j*} for which the values of the contextual parameters of the m observations of $\mathbf{X}$ assume the biggest sum of r_{ij*}, $i = 1,\ldots, n$, $i \notin J$, that is

$$j^* = \arg\max_{j \in J} \left\{ \sum_{i=1, i \notin J}^{n-|J|} r_{ij} \right\}$$

3. We estimate the missing value of y_{j*}^{m+1} of X_{j*}, i.e., $y_{j*}^{m+1} = f(x^{m+1} \mid \theta_{j*}')$, where θ_{j*}' is the parameter vector for WLS for regression of X on X_{j*}.
4. We repeat Step 3 for all $j \in J$ and in each step we expand $\mathbf{X}_D^{m+1}$ with the extrapolated value y_{j*}^{m+1}.

4.2 On the use of Recursive Probabilistic PCA

We can estimate the missing values of $\mathbf{X}$ using the statistical method of PCA. Let $\mathbf{Z} = [Z_1, \ldots, Z_n]$ be the *standard score* of vector $\mathbf{X}$, i.e.,

$$Z_j = \frac{X_j - b_j}{\sigma_j} \quad (2)$$

for $j = 1, \ldots, n$, where b_j and σ_j are respectively the mean and standard deviation of parameter X_j. We discard these rows and columns from the sample correlation coefficient matrix $\mathbf{R} = [r_{ij}] \in \Re^{n \times n}$ of $\mathbf{Z}$ which are $r_{ik} = r_{ki} < \epsilon$, $\forall\ i = 1, \ldots, n$. The matrix that remains is the sample correlation coefficient matrix $\mathbf{R}_C = [r_{c,ij}] \in \Re^{n_c \times n_c}$ of parameters which are the most correlated ones, $n_c \leq n$. To simplify the notation, we consider that the correlated parameters are the first n_c of these. Let $\mathbf{Z}_C = [Z_1, \ldots, Z_{n_c}]$ be the n_c - dimensional vector of n_c correlated contextual parameters and $\mathbf{X}_C = [X_1, \ldots, X_{n_c}]$ the respective of the real observations. PCA is an orthogonal linear transformation that transforms the n_c correlated parameters into n_c new linear independent variables called *principal components* (PC). PC are the directions of $\Re^{n_c}$ that coincide with the eigenvectors $\mathbf{w}_i$, $i = 1, \ldots, n_c$ of matrix $\mathbf{R}_C$. The elements w_{ij} of eigenvector $\mathbf{w}_i$ are the coefficients of linear combination of parameters that define the direction of i^{th} PC.

Let $\lambda_1, \ldots, \lambda_{n_c}$ be the eigenvalues of matrix $\mathbf{R}_C$, where $\lambda_1 > \ldots > \lambda_{n_c}$. Let $\mathbf{\Lambda}_C = \text{diag}(\lambda_1, \ldots, \lambda_{n_c})$ and $\mathbf{W}_C = [\mathbf{w}_1, \ldots, \mathbf{w}_{n_c}] \in \Re^{n_c \times n_c}$, where $\mathbf{w}_i$ is the eigenvector corresponding to eigenvalue λ_i. Hence,

$$\mathbf{R}_C = \mathbf{W}_C \cdot \mathbf{\Lambda}_C \cdot \mathbf{W}_C^{\mathsf{T}} \quad (3).$$

We assume that $\mathbf{X}_C$ follows a multivariate normal distribution. Additionally, we assume that we have already constructed a PCA model and that the new $(m+1)^{\text{th}}$ vector $\mathbf{X}_C^{m+1}$ does not affect considerably the values of elements of matrix $\mathbf{R}_C$. We want to estimate the contextual missing values that occur in the $\mathbf{X}_C^{m+1}$. The missing data can be assumed to be the first elements of the context vector without loss of generality. We denote the missing and the observed parameters with $\mathbf{X}_C^{\#,m+1}$ and $\mathbf{X}_C^{*,m+1}$ respectively, so the vectors $\mathbf{X}_C^{m+1}$ and $\mathbf{Z}_C^{m+1}$ can be partitioned as:

$$\mathbf{X}_C^{m+1\mathsf{T}} = [\mathbf{X}_C^{\#,m+1\mathsf{T}}\ \mathbf{X}_C^{*,m+1\mathsf{T}}]$$

and

$$\mathbf{Z}_C^{m+1\mathsf{T}} = [\mathbf{Z}_C^{\#,m+1\mathsf{T}}\ \mathbf{Z}_C^{*,m+1\mathsf{T}}].$$

Correspondingly, the $\mathbf{W}_C$ matrix can be partitioned as

$$\mathbf{W}_C^{\mathsf{T}} = [\mathbf{W}_C^{\#\mathsf{T}}\ \mathbf{W}_C^{*\mathsf{T}}].$$

We substitute the missing values $\mathbf{Z}_C^{\#,m+1}$ with the expected values from the conditional normal distribution given the present $\mathbf{Z}_C^{*,m+1}$ vector and the current estimate of correlation matrix:

$$\hat{\mathbf{Z}}_C^{\#,m+1} = E\{\, \mathbf{Z}_C^{\#,m+1} \mid \mathbf{Z}_C^{*,m+1}, \mathbf{R}_C \,\}$$

Nelson P. et al. [20] stated that the conditional expectation of the missing measurements is given by:

$$\hat{\mathbf{Z}}_C^{\#,m+1} = \mathbf{W}_C^{\#} \cdot \mathbf{\Lambda}_C \cdot \mathbf{W}_C^{*\mathsf{T}} \cdot (\mathbf{W}_C^{*} \cdot \mathbf{\Lambda}_C \cdot \mathbf{W}_C^{*\mathsf{T}}) \cdot \mathbf{Z}_C^{*,m+1} \quad (4)$$

The above values are those that the expectation – maximization (EM) ([22]) algorithm calculate in expectation step. This algorithm can also be used when we construct the PCA model to handle the missing values.

From Equation (2) we can calculate the estimate $\hat{\mathbf{X}}_C^{\#,m+1}$ of $\mathbf{X}_C^{\#,m+1}$ vector. A problem that we must solve in the proposed monitoring system is the renewal of model in processes that change over. Our model is based on m last $\mathbf{X}_C^k$ vectors of measurements. We renew some or all parameters of model, depending on the application. As long as we focus on critical applications, an efficient way is to compute recursively the new values of model's parameters when a new vector of measurements comes. Tien D. et al. ([23]) concluded that when we apply the conventional PCA using a static model we reduce the false alarms if we convert the measurements in standard score using the mean and standard deviation of sample instead of corresponding values that we used on model's construction. We counter the problem of unknown mean and standard deviation of the sample by computing them from a moving window that contains the m last vectors, which represent better the current state of sample.

The computations of new mean vector $\mathbf{b}_C^{m+1} \in \Re^{n_c \times n_c}$ and standard deviation σ_j^{m+1} after the reception of the $(m+1)^{\text{th}}$ measurements vector are necessary at any case. We can compute recursively these parameters of model from previous - historical - sample's values $\mathbf{b}_C^m$ and σ_j^m, respectively, by next recursive equations:

$$\mathbf{b}_C^{m+1} = \mathbf{b}_C^m + \frac{\mathbf{X}_C^{m+1} - \mathbf{X}_C^1}{m} \quad (5)$$

and

$$\left(\sigma_j^{m+1}\right)^2 = \left(\sigma_j^m\right)^2 + \frac{(X_j^{m+1})^2 - (X_j^1)^2 - m \cdot ((b_j^{m+1})^2 - (b_j^m)^2)}{m-1} \quad (6)$$

The computational cost is constant for every one of n_C parameters and this is important for systems with bounded computational power. Especially, medical data through pervasive sensors may have parameters that remain constant for long intervals of time, like heart's pulse and data of oxymeter sensor, when the patient acts firmly. These parameters have zero sample standard deviation. Thus, Equation (2) gives indefinable results

(NaNs). For this reason, outliers replace missing values. To deal this, we can use a univariate method to substitute the missing values or we can assume that these parameters are also missing and compute the missing value by contribution of remaining parameters. However, the former approach biases the parameter and the latter gives less accurate values. If we deem that the model of process must be updated, we can compute recursively the new correlation matrix $\mathbf{R}_C^{m+1}$ from the previous matrix $\mathbf{R}_C^m$ using the equation:

$$\mathbf{\Sigma}_C^{m+1} \cdot \mathbf{R}_C^{m+1} \cdot \mathbf{\Sigma}_C^{m+1} = \mathbf{\Sigma}_C^m \cdot \mathbf{R}_C^m \cdot \mathbf{\Sigma}_C^m + \mathbf{A} \quad (7)$$

with

$$\mathbf{A} = \frac{\mathbf{X}_C^{m+1}\mathbf{X}_C^{m+1\top} - \mathbf{X}_C^1\mathbf{X}_C^{1\top}}{m-1} - \frac{m(\mathbf{b}_C^{m+1}\mathbf{b}_C^{m+1\top} - \mathbf{b}_C^m\mathbf{b}_C^{m\top})}{m-1}$$

where $\mathbf{\Sigma}_C^m = \mathrm{diag}(\sigma_1^{m+1}, \ldots, \sigma_{n_C}^{m+1})$ and $\mathbf{\Sigma}_C^m = \mathrm{diag}(\sigma_1^m, \ldots, \sigma_{n_C}^m)$.

Furthermore, we compute recursively the eigenpairs $\{\mathbf{wi}, \lambda_i\}$ using the algorithm proposed by Bunch J.R. et al. ([24]).

5. PERFORMANCE ASSESSMENT

To evaluate the substitution systems, we gathered medical data that consisted of nine parameters. Six of them are accelerations that are produced from two sensor nodes. ECG, arterial pressure and blood oxygen saturation completed the 9-dimensional context vector of 9 contextual medical parameters. We describe the technical characteristics of nodes of biosensor network that produced the measurements in the following paragraphs.

The Sentilla Perk [13] sensor kit has been utilized in our system. The latter contains two 2.4 GHz wireless data transceivers (nodes, see fig. 2) using the IEEE 802.15.4 (ZigBee) protocol. It also includes a USB port for interface with a personal computer acting as the monitoring unit. Each node has a low-power, low-voltage MCU (MicroController Unit), one 3D Accelerometer for X, Y and Z axis and additional analog and digital input pins for adding more sensors. The Perk nodes are provided in a plastic robust small-sized enclosure (6x3x1.5cm) making them more suitable for placing on patient's body and tolerating falls.

(a)

(b) (c)

Figure 4. On-body sensor devices utilized for patient data acquisition: a) The Sentilla Perk node ([13]) containing a 3D accelerometer that can be attached on user and send motion data through the ZigBee wireless protocol. The plastic enclosure can protect the node from falls and makes it more suitable for carrying it on patient's body, b) sentilla node connected to ECG board [[14], c) Wireless pulse oxymeter [15]

Two Perk nodes can be placed on patient's body. Preferable positions are close to user's chest and user's belt or lower at user's foot. The latter positions have proven based on conducted experiments to be appropriate for distinguishing rapid acceleration on one of the three axis that is generated during a fall. The nodes have two analog input interfaces that can be exploited for connecting biosignal sensors like the ECG circuit in Figure 4 developed according to [14]. Appropriate J2ME code is developed and deployed on the nodes for reading the accelerometer values and analog values from the attached sensors and transmitting them wirelessly to the monitoring unit. At the latter a Java application built using the Sentilla IDE [13] receives the patient data and performs further processing as described in the following sections. The X, Y and Z acceleration values from both sensors are interlaced. The ECG signal can be further processed in order to acquire information like salient complexes (i.e. QRS complex), and detect specific patterns like arrhythmias, etc.

An additional wireless pulse oxymeter sensor [15] has been used to provide more information related to the patient's physiological state. Arterial pressure and blood oxygen saturation level are wirelessly transmitted to the monitoring unit from the device. The device has an embedded sound alarm mechanism that can notify caregivers in case predefined thresholds for arterial pressure and oxygen levels are exceeded.

The above nodes have different sampling rates. We selected 521 vectors of measurements by rate of 2 samples per second using network about 4.5 min. We calculated the correlation matrix of sample and we used it to build the model. We put randomly missing values in the remaining vectors and reconstructed the failures. In Fig.5 we show one acceleration signal and the signals of the ECG and of arterial pressure.

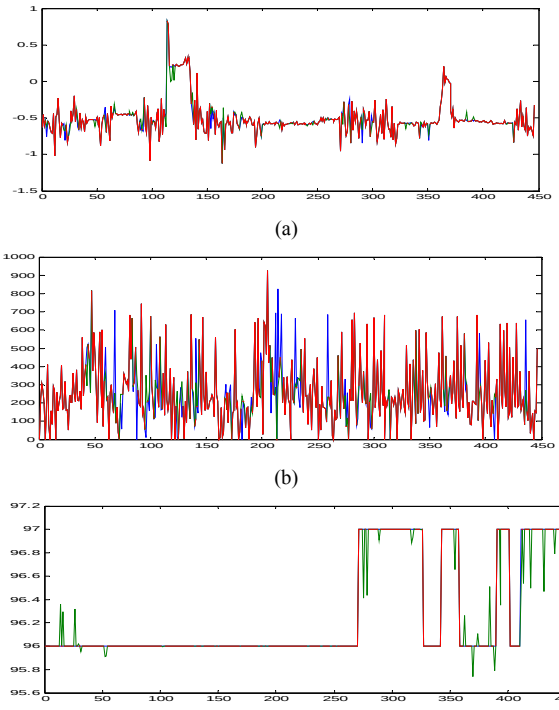
Because of the different nature of parameters, we used the sum of squares of reconstruction errors (RE) $e_j = \hat{\mathbf{Z}}_j^{\#,m+1} - \mathbf{Z}_j^{\#,m+1}$ of parameters in standard scores as metric. Using the mean and standard deviation of samples, whose sizes were m, $25 \le m \le 50$, we calculated the RE. We conclude that metric is reduced when the m grows (see Fig. 6.a). Following, we calculated correlation matrices and built PCA models based on different m numbers of vectors, $25 \le m \le 50$. We recalculated the estimates of signal based on these models keeping the m recent vectors. At all events we calculated again the above metric. The reduction of metric appeared again when the m grew (see Fig.6.b). Moreover, we saw that the model generated by large number of vectors gave better estimates (see Fig.6.a).

We considered the reconstruction of each parameter when this parameter is systematically missing. As it is expected, the reconstructions are better when the parameters are strongly correlated and their standard deviations are important. Thus, every parameter has its contribution to sum of squares of RE.

We created more different erroneous signals samples by the random putting of the missing values. The appearance's rates of the missing values of these samples were different. We executed

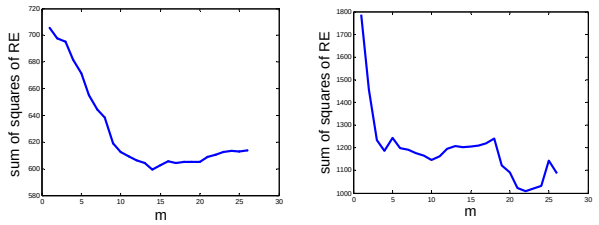
the RPPCA algorithm using these samples. Thus, we conclude that the sum of squares of RE converges when the rate of missing values is large. This occurs because the missing values are statistically distributed equally among the parameters (see Fig. 7).

The quality of the estimates is depended on the type of the missing parameter. The RE is less for the most correlated parameters. Parameters with a small standard deviation are deviating from the multivariate normal distribution. These parameters are not be estimated by a mechanism that uses PCA, because this is based on the standard deviation of measurements. Instead, we can use a simple imputation, like the substitution by previous value to estimate the missing values of these. Moreover, it is better to eliminate these parameters from the context vector $\mathbf{X}_C$, during the intervals that are constant.



(a) (b) (c)

Figure 5. The original signal is blue, the reconstructed is green and the inaccurate is red. (a) One acceleration signal. (b) ECG signal (c) Arterial pressure signal



(a) (b)

Figure 6. The sum of squares of RE using (a.) all available and (b.) *m* vectors to construct PCA model with *m* size of moving window.

In addition, we examine the behavior of the FIS in the missing value substitution problem. We assess the capability of the FIS to identify any missing value for a sensor stream with correlated contextual parameters. Specifically, we measure the identification error each time a missing value occurs, i.e., $e(t) = \|\mathbf{X}(t) - \mathbf{X'}(t)\|$. That is, the context $\mathbf{X}$ at time t is the vector $\mathbf{X}(t) = [X_1(t), X_2(t), X_3(t)]$ and $X_3(t)$ is strongly correlated with $X_1(t)$ and $X_2(t)$ (we set correlation threshold $\varepsilon = 0.8$ – see Section 4.1). The missed context $\mathbf{X'}(t)$ derives if we miss the value of X_3 every two samples, i.e., frequency of missing value is 0.5Hz providing that we obtain one sample every second. In the following experiment we correlate the X and Y acceleration values from sensors with the Z acceleration and we require that the FIS is able to identify and replace any missing value from the Z acceleration. Fig. 8 depicts the $e(t)$ of the identification of the missing context $\mathbf{X'}(t)$. In addition, we evaluate the identification system with $\mathbf{X}(t) = [X_1(t), X_2(t), X_3(t)]$ where there is correlation between the $X = X_1(t)$ and $Y = X_2(t)$ acceleration values from sensors with the ECG signal X_3. As shown in Fig. 9, the FIS identifies such correlation and replaces the missed ECG signal satisfyingly.

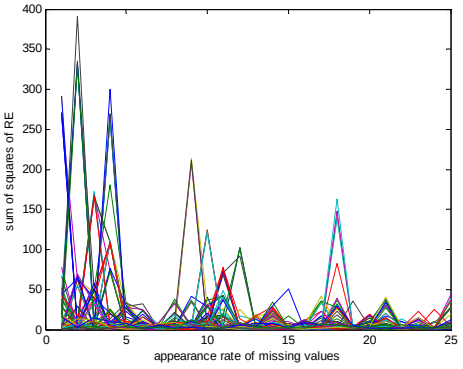


Figure 7. The sum of squares of RE of 100 samples in regard of appearance rate of missing values.

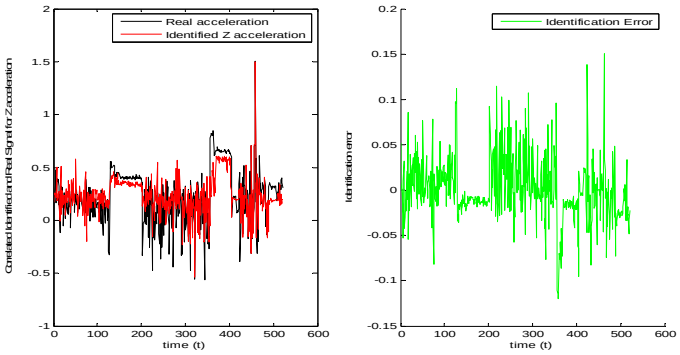


Figure 8. The identified Z acceleration and the corresponding identification error when adopting the FIS.

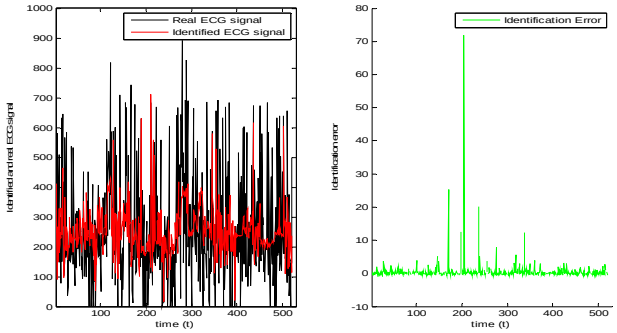


Figure 9. The identified ECG signal and the corresponding identification error when adopting the FIS.

6. CONCLUSIONS

In this paper, we proposed two methods of substitution of the missing values derived from data streams of wireless bio-sensor networks. These methods are based on linear relations among the contextual parameters that are used to monitoring patients state. One of these, the RPPCA has low computational cost in some cases. In addition the FIS replaces any identified correlated missing value but with more computational const than the RPPCA. Moreover, the FIS assumes that the missing values have at least some correlation with the observations. We discussed a problem that derived from the nature of some medical parameters like heart's pulse. We are currently implementing the functionality of the described methods in wireless sensor networks on some areas, like environment monitoring. We are planning to extend these methods to detect other types of failures, like outliers that are not evident. This is important in health care applications because it can reduce the false alarms or detect possibly the aberrations of nature functions.

7. REFERENCES

[1] Juan M. Corchado, Javier Bajo, Yanira de Paz, Dante I. Tapia: Intelligent Environment for monitoring Alzheimer patients, agent technology for healthcare, to be published in Decision Support Systems, article available online at www.sciencedirect.com

[2] Moushumi Sharmin, Shameem Ahmed, Ahamed S.I., Haque M.M., Khan A.J.: Healthcare aide: towards a virtual assistant for doctors using pervasive middleware, In Proc. of Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops (2006) 6-12.

[3] Paganelli F., Spinicci E., Mamelli A., Bernazzani R., Barone P.: ERMHAN: A multi-channel context-aware platform to support mobile caregivers in continuous care networks, In Proc. of IEEE International Conference in Pervasive Technologies (2007) 355-360.

[4] Scott Mitchell, Mark D. Spiteri, John Bates and George Coulouris: Context-Aware Multimedia Computing in the Intelligent Hospital, In Proc. SIGOPS EW2000, the Ninth ACM SIGOPS European Workshop (2000).

[5] Gouaux F., Simon-Chautemps L., Adami S., Arzi M., Assanelli D., Fayn J., Forlini M.C., Malossi C., Martinez A., Placide J., Ziliani G.L., Rubel P.: Smart devices for the early detection and interpretation of cardiological syndromes, In 4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine (2003) 291-294.

[6] Youngho Jeen, Jiyoung-Kim, Jaeil Park, Peom Park: Design and implementation of the Smart Healthcare Frame Based on Pervasive Computing Technology, In The 9th International Conference on Advanced Communication Technology (2007) 349-352.

[7] Jannett T.C., Prashanth S., Mishra S., Ved V., Mangalvedhekar A., Deshpande J.: An intelligent telemedicine system for remote spirometric monitoring In Proceedings of the Thirty-Fourth Southeastern Symposium on System Theory (2002) 53-56.

[8] Dolgov A.B., Zane R.: Low-Power Wireless Medical Sensor Platform, In 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2006) 2067-2070.

[9] C. Doukas, I. Maglogiannis, G. Kormentzas: Advanced Telemedicine Services through Context-aware Medical Networks, In International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine (2006).

[10] C. Doukas, T. Pliakas, I. Maglogiannis: Advanced Scalable Medical Video Transmission based on H.264 Temporal and Spatial Compression, Presented at 2007 IEEE Africon Conference (2007).

[11] David Malan, Thaddeus Fulford-Jones, Matt Welsh, and Steve Moulton: CodeBlue: An Ad Hoc Sensor Network Infrastructure for Emergency Medical Care, International Workshop on Wearable and Implantable Body Sensor Networks (2004).

[12] Sokwoo Rhee, Boo-Ho Yang, Kuowei Chang and Hamhiko H. Asada: The Ring Sensor: a New Ambulatory Wearable Sensor for Twenty-Four Hour Patient Monitoring, In Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 20 4 (1998) 1906-1909.

[13] The Sentilla Perk Pervasive Computing Kit, http://www.sentilla.com/perk.html

[14] T.J. Sullivan, S.R. Deiss, G. Cauwenberghs, "A Low-Noise, Non-Contact EEG/ECG Sensor", Biomedical Circuits and Systems Conference, 2007. BIOCAS 2007. IEEE. 27/12/2007; DOI: 10.1109/BIOCAS.2007.4463332.

[15] Wireless Pulse Oxymeter by FaceLake®, more information at http://www.facelake.com/cms50e.html

[16] Little R., Rubin D.: "Statistical Analysis with missing data" (A Wiley-Interscience Publication J. Wiley & sons, 1987)

[17] Jollifee, I.T.: "Principal Component Analysis" (Springer – Verlag, New York Inc., 2002)

[18] Wasito I., Mirkin B., "Nearest neighbour approach in the least-squares data imputation algorithms", Information Sciences 169 (2005) 1–25

[19] Dunia R., Qin S. J., Edgar T. F., McAvoy T. J., "Identification of Faulty Sensors Using Principal Component Analysis", AIChE Journal,October, vol. 42, no 10 (1996), pp. 2797-2812.

[20] Nelson P., Taylor P., McGregor J., "Missing data methods in PCA and PLS: Score calculations with incomplete observations", Chemometrics and Intelligent Laboratory Systems, vol. 35 (1996), pp. 45–65

[21] Arteaga F., Ferrer A., "Dealing with missing data in MSPC: several methods, different interpretations, some examples", Journal of chemometrics, vol. 16 (2002), pp. 408–418

[22] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the E-M algorithm. Journal of the Royal Statistical Society, Series B, 39(1),pp. 1–38

[23] Tien D., Lim K.–W., Jun L. "Comparative Study of PCA Approaches in Process Monitoring and Fault Detection", The30th Annual Conference of the IEEE Industrial Electronics Society, November 2004, Busan, Korea, 2594–2599.

[24] Bunch J.R., Nielsen C.P., Sorensen D.C., "Rank-one modification of the symmetric eigenproblem", Numerische Mathematik 31 (1978), pp. 31–48

[25] K. Passino and S. Yurkovich, *Fuzzy Control*, Addison Wesley Longman, Menlo Park, CA, 1998