A Novel Leak Detection Approach in Water Distribution Networks

Radhia Fezai  
and Kais Bouzrara

Majdi Mansouri, Hazem Nounou  
and Mohamed Nounou

University of Monastir, Tunisia  
Texas A&M University at Qatar, Qatar

Vicenç Puig

Polytechnic University of Catalonia, Barcelona, Spain

Abstract—This paper proposes a novel leak monitoring framework aims to improve the operation of water distribution network (WDN). To do that, an online statistical hypothesis test based leak detection is proposed. The main advantages of the developed method are first to deal with the higher required computational time for detecting leaks and then, to update the KPCA model according to the dynamic change of the process. Thus, this can be performed to massive and online datasets. Simulation results obtained from simulated WDN data demonstrate the effectiveness of the proposed technique.

Index Terms—Leak detection, water distribution networks, kernel principal component analysis, online reduced kernel generalized likelihood ratio test, exponentially weighted moving average.

I. INTRODUCTION

This paper allows to evaluate the state of the water distribution network (WDN) (to decide if a leak exists). The idea behind leak detection water distribution systems is to maintain disinfection levels, pressure and reduce water loss are equally important. To address the leak detection purposes, several approaches have been developed including principal component analysis (PCA) [1], [2], nonlinear PCA (NPCA) [3], Multi-Regional PCA (MRPCA) [4], probabilistic PCA (PPCA) [5], [6] and attribute PCA (APCA)[7]. However, most industrial systems are nonlinear. Thus, kernel PCA (KPCA) [8], [9], [10], [8] is developed to extend PCA to nonlinear systems. KPCA method consists to project the input data into a high-dimensional space using nonlinear kernel functions. The main advantage of KPCA is that it does not involve non-linear optimization since it can efficiently determine principal component in the high-dimensional space by means of integral operators and nonlinear mapping. In addition, it requires only the determination of eigenvalue problem. Due to these merits, KPCA has proven successful in the detection sector [11]. However, the classical KPCA technique is not appropriate to deal with dynamic systems since they are basically non-adaptive procedures. This technique requires to keep the training data set and then use the KPCA model for detection purposes. For large number of samples, the detection using KPCA method imposes a high computational cost. Then, KPCA is not adequate for real processes that are prone to drifting off due to various phenomena since the system may be subject to significant changes. This drift can change the relationship between process variables and can lead to poor detection results.

To deal with changing process conditions and high computational cost, we propose to update the KPCA model and decrease the number of initial samples. Thus, a reduced KPCA method is developed[12], [13]. This method aims to select a new reduced data set, so-called dictionary, to construct the reduced KPCA model. The resulting dictionary formed by linearly independent kernel functions. The developed technique is applied in an online phase to construct the reduced model and it takes into account the dynamic behavior of the process by changing the model structure. Then, it uses the new KPCA model for fault detection purposes. To detect the presence of fault, the most used indices [14] are the Hotelling’s $T^2$ statistic [15] which is performed to monitor the principal component subspace in the feature space, the squared prediction error (SPE) [16] and kernel generalized likelihood ratio test (KGLRT) that monitor the residual subspace[17], [18]. In [19], KGLRT index has been shown to outperform other fault detection statistics. However, the main disadvantage of KGLRT-based monitoring technique is the time-invariance of its control limit and KGLRT is not able noise in practical measurements. To further enhance the fault detection abilities of the KGLRT statistic, we propose to use the exponentially weighted online reduced Kernel generalized likelihood ratio test (EW-ORKGLRT) technique.

The idea behind the proposed leak detection technique is to use the reduced KPCA technique, which minimizes the computational cost of the classical KPCA model, to update the model and then, the reduced model is introduced to EW-ORKGLRT chart to perform the detection phase in the leak monitoring prob-
lem. The advantages of the developed technique are that the EW-ORKGLRT statistic takes into account the information provided using the current and previous sample by giving significant importance to the more recent information. Moreover, the computational time of storing and evaluating the EW-ORKGLRT statistic is decreased significantly by using a reduced data set. Also, the adaptive control chart according to the dynamic process change improves the performance of detection by reducing the false alarm rate and ameliorates the good detection rate.

The remainder of the paper is organized as follows. Section 2 presents KPCA model based detection technique. In Section 3, the improved leak detection approach using EW-ORKGLRT is developed. Then, in Section 4, the leak detection performance is evaluated using a simulated WDN data. At the end, the conclusions are presented in Section 5.

II. KERNEL GLRT BASED ON KPCA MODEL

A. Description of KPCA model

KPCA technique consists to map the initial data matrix \( X = [x_1 \ldots x_N]^T \), \( X \) contains \( N \) samples and \( m \) variables, into a feature space using a nonlinear mapping function \( \Phi \) and then a linear PCA is applied in feature space. The mapped data matrix is denoted as \( \Phi(X) = [\Phi(x_1) \ldots \Phi(x_N)]^T \in \mathbb{R}^{N \times h} \) with \( h \) is a large value representing the dimension in feature space. It is assumed that \( \Phi(X) \) has been mean-centered. The covariance matrix is determined in the higher-dimensional spaces is given by [20]

\[
Q = \frac{1}{N-1} \Phi^T(X) \Phi(X) = \frac{1}{N-1} \sum_{i=1}^{N} \Phi(x_i) \Phi^T(x_i)
\]  

Then, the main idea of KPCA is to seek the eigenvalue \( \mu \) and their corresponding eigenvector \( v \) by solving the eigenvector problem,

\[
Qv = \mu v
\]  

To resolve equation 2, it is required to compute the inner product of two kernel functions by using a kernel function,

\[
k(x_i, x_j) = \Phi^T(x_i) \Phi(x_j)
\]  

Radial basis kernel is the most used kernel function,

\[
k(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{\epsilon}\right)
\]  

where \( \epsilon \) is the kernel parameter.

Denoting \( K = \Phi(X) \Phi^T(X) \) the kernel matrix. According to equation 1, equation 2 can be calculated as,

\[
(N-1)\mu v = \Phi^T(X) \Phi(X) v
\]  

Now, by multiplying with \( \Phi(X) \) from the left of both sides in equation 5, the following expression is obtained

\[
\lambda \Phi(X) v = \Phi(X) \Phi^T(X) \Phi(X) v
\]  

where \( \lambda = (N-1)\mu \).

The eigenvector problem can be expressed as,

\[
\lambda \Phi(X) v = K \Phi(X) v
\]  

Then, define

\[
\alpha = \Phi(X) v
\]  

In this regard, equation 7 is transformed into,

\[
\lambda \alpha = K \alpha
\]  

According to equation 9, \( \lambda \) identifies the eigenvalue corresponding to the eigenvector \( \alpha \) of the kernel matrix \( K \). The eigenvector \( v \) is given by,

\[
v = \lambda^{-1} \Phi^T(X) \alpha
\]  

Defining \( P_f = \begin{bmatrix} v_1 & \cdots & v_\ell \end{bmatrix} \) the matrix of \( \ell \) principal eigenvectors. From equation 10, \( P_f \) is defined as,

\[
P_f = \begin{bmatrix} \frac{1}{\lambda_1} \Phi^T(X) \alpha_1 & \cdots & \frac{1}{\lambda_\ell} \Phi^T(X) \alpha_\ell \end{bmatrix} = \Phi^T(X) P \Lambda^{-1}
\]  

where \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_\ell) \) and \( P = \begin{bmatrix} \alpha_1 & \cdots & \alpha_\ell \end{bmatrix} \) and are the \( \ell \) largest eigenvalues and principal eigenvectors of the matrix \( K \), respectively.

The number of kernel principal components (KPCs) \( \ell \) is determined using the cumulative percent variance (CPV) criterion.

For the test measurement \( x \) and its projected vector \( \Phi(x) \), its projections on the eigenvectors, also called the kernel components, are computed as,

\[
t = P_f^T \phi(x)
\]  

According to equation 11, the vector \( t \) can be expressed as,

\[
t = \Lambda^{-1} P \Phi(X) \Phi(x) = \Lambda^{-1} P k(x)
\]  

where \( k(x) = [k(x_1, x)(x_N, x)]^T \).
**B. Description of kernel GLRT based detection chart**

The exponentially weighted kernel GLRT (EW-KGLRT) detection chart merges the advantages of the exponentially weighted moving average (EWMA) and kernel GLRT charts. It aims to determine a new KGLRT index that includes current and previous samples in a decreasing exponential mode by providing more weight to the more recent sample. The KGLRT chart is one of the most used techniques for solving composite hypotheses testing problems by maximizing the likelihood ratio function over all faults [21]. The main idea begins with one N-dimensional vector $\Phi(x)$, the hypothesis testing problem is given by, [22]:

$$H_0 : \Phi(x) = \omega \quad \text{(null hypothesis)}$$

$$H_1 : \Phi(x) = P\theta + \omega \quad \text{(alternative hypothesis)}$$  

where $\omega$ represents a white noise following a normal distribution $N(0, \sigma^2 I)$, with $I$ is an identity matrix, $P$ is an orthogonal matrix $(P^T = I)$ and $\theta$ is the mean vector.

The likelihood ratio test for a new kernel function $\Phi(x)$ method chooses between $H_0$ and $H_1$ [22] as:

$$\rho = \frac{f_1(\Phi(x) \mid H_1)}{f_0(\Phi(x) \mid H_0)} \leq \gamma$$  

where $\gamma$ is the threshold value of KGLRT statistic.

$$f_0(\Phi(x) \mid H_0) = \frac{1}{(2\pi \sigma_0^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_0^2}||\omega_0||^2\right)$$

$$f_1(\Phi(x) \mid H_1) = \frac{1}{(2\pi \sigma_1^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_1^2}||\omega_1||^2\right)$$

where $\omega_0 = \Phi(x)$ and $\omega_1 = \Phi(x) - P\theta$. Here, the KGLRT chart changes the unknown parameter $\theta$, $\sigma_0$ and $\sigma_1$ by its maximum likelihood estimate $\hat{\theta}$, $\hat{\sigma}_0$ and $\hat{\sigma}_1$. The maximum likelihood estimate of $\theta$ is identical to the least square estimate of $\omega_1$ [17].

$$\hat{\omega}_0 = \Phi(x)$$

$$\hat{\omega}_1 = \Phi(x) - P\hat{\theta} = (I - C_f)\Phi(x)$$

where $C_f = P_f P_f^T$.

The estimate of $\hat{\sigma}_0$ and $\hat{\sigma}_1$ are given by,

$$\hat{\sigma}_0 = \frac{1}{N}||\hat{\omega}_0||^2$$

$$\hat{\sigma}_1 = \frac{1}{N}||\hat{\omega}_1||^2$$

Inserting the maximum likelihood estimates of the parameters (equation 19) into equation 15 and taking $N/2$ root, KGLRT is defined as,

$$\begin{align*}
\mathbb{K}_G &= \frac{||\omega_0||^2}{||\omega_1||^2} = \frac{\Phi^T(x)I\Phi(x)}{\Phi^T(x)(I - C_f)\Phi(x)} \\
&= \frac{\Phi^T(x)\Phi(x)}{\Phi^T(x)\Phi(x) - \Phi^T(x)P_fP_f^T\Phi(x)}
\end{align*}$$

Then, from equation 11, the KGLRT statistic can be computed as,

$$\begin{align*}
\mathbb{K}_G &= \frac{1}{1 - \Phi^T(x)P\Lambda^{-1}P^T\Phi(x)} \\
&= \frac{1}{1 - k^T(x)P\Lambda^{-2}P^Tk(x)}
\end{align*}$$

To improve the performance of the classical KGLRT test, a new EW-KGLRT index that combines the advantages of KGLRT chart with the exponentially weighted filter is proposed. The exponentially weighted-KGLRT statistic gives effective and fast detection while conserving a low false alarm rate and higher good detection rate. The EW-KGLRT chart $\mathbb{EKG}$ can be computed as:

$$\mathbb{EKG} = \varpi\mathbb{K}_G + (1 - \varpi)\mathbb{EKG}$$

where $\varpi$ define the smoothing parameter of the EWMA filter. To determine the threshold for the filtered KGLRT, its distribution should be determined. The new filtered KGLRT statistic follows a Chi-square distribution $\chi^2$ since KGLRT distributed according to a Chi-square distribution [23]. Thus, the threshold for $\mathbb{EKG}$ statistic is determined using the $\chi^2$-distribution and it is computed as, [17]:

$$\mathbb{EKG}_\alpha = g_{\mathbb{EKG}}\chi^2_{\mathbb{EKG},\alpha}$$

where $g_{\mathbb{EKG}} = \frac{b_{\mathbb{EKG}}}{2a_{\mathbb{EKG}}}$ and $h_{\mathbb{EKG}} = \frac{2a_{\mathbb{EKG}}}{b_{\mathbb{EKG}}}$, with $a$ and $b$ are the mean and variance of the $\mathbb{EKG}$. The $\mathbb{EKG}$ statistic indicates the the existence of a fault if it satisfies the following condition

$$\mathbb{EKG}(x) > \mathbb{EKG}_\alpha$$

**III. Description of the developed leak detection technique**

**A. Online leak detection based EW-ORKGLRT statistic**

In this study, an online leak detection based EW-ORKGLRT statistic algorithm is derived. In this algorithm, there are two possible cases: keep the dictionary unchanged or expand it with the new kernel function.
A different dictionary will yield a different Gram matrix, eigenvector and detection EW-ORKGLRT statistic. These parameters are updated only due to the dictionary change. Now, consider the case when the sample \( x_{k+1} \) is leaky. It implies that EW-ORKGLRT statistic verifies this relation; \( \text{EKG}_{k+1} \leq \text{EKG}_{\alpha,k} \) and the sample is not included in the dictionary. Thus, the dictionary remains unchanged. For leaky data \( \text{EKG}(x_{k+1}) \leq \text{EKG}_\alpha \), the dictionary keeps the same observations at instant \( k \). However, for normal data \( \text{EKG}(x_{k+1}) \geq \text{EKG}_\alpha \), two cases can arise:

In the first case, the dictionary is unchanged and the kernel function \( \Phi(x_{k+1}) \) is removed from the dictionary.

\[
D_{k+1} = \{D_k\} \tag{25}
\]

The data is not added to the dictionary if

\[
\varepsilon_{k+1} = k(x_{k+1}, x_k) - k'(x_{k+1})^T \beta < \nu \tag{26}
\]

This means that the kernel function \( \Phi(x_{k+1}) \) may be approximated by a linear combination the other elements of the dictionary \( D_k \).

The vector \( k'(x_{k+1}) \) is updated as,

\[
k'(x_{k+1}) = [k(x_{w_1}, x_{k+1}) \cdots k(x_{w_r}, x_{k+1})]^T \tag{27}
\]

The vector \( \beta = \begin{bmatrix} \beta_1 \cdots \beta_r \end{bmatrix}^T \) with the dictionary unchanged can be updated as

\[
\beta = (K'^r)^{-1} k'(x_{k+1}) \tag{28}
\]

where \( K'^{r+1} \in \mathbb{R}^{r \times r} \) is the reduced Gram matrix obtained from the dictionary \( D_{k+1} \) that contains the elements \( k(x_i, x_j) \), \( i, j = w_1, \ldots, w_r \).

From equation 10, the eigenvector \( V'_{k+1} \) can be updated as,

\[
u'_{k+1} = \sum_{i=1}^{r} \alpha'_{k+1,i} \Phi(x_{w_i}), \quad k = 1, \ldots, N \tag{29}
\]

For the second case when the new kernel function \( \Phi(x_{k+1}) \) is significantly different from the previously selected elements of the dictionary, it is added into the dictionary and the new dictionary \( D_{k+1} \) becomes,

\[
D_{k+1} = \{D_k, \Phi(x_{k+1})\} \tag{30}
\]

In this case, the dimensionality of the Gram matrix increases. Thus, the new Gram matrix \( K'^{r+1} \in \mathbb{R}^{r+1 \times r+1} \) is updated as

\[
K'^{r+1}_{k+1} = \begin{bmatrix} K'^r_k & k'(x_{k+1}) \\ k'(x_{k+1})^T & k(x_{k+1}, x_{k+1}) \end{bmatrix} \tag{31}
\]

To avoid the problem of the higher computational complexity, the inverse of the kernel matrix \( (K'^{-1})_{k+1} \) is calculated iteratively from the Woodbury matrix identity [14]:

\[
K'^{-1}_{k+1} = \begin{bmatrix} (K'^{-1}_k)^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\varepsilon_{k+1}} \begin{bmatrix} \alpha'(x_{k+1})^{-1} k'(x_{k+1})^T & -k'(x_{k+1})^T (K'^{-1}_k)^{-1} \\ -k'(x_{k+1}) (K'^{-1}_k)^{-1} & \frac{1}{\varepsilon_{k+1}} \begin{bmatrix} -\beta_k \beta_k \end{bmatrix} \end{bmatrix} \tag{32}
\]

The updating of the kernel vector with the new kernel function inserted into the dictionary becomes

\[
k'(x_{k+1}) = [k(x_{w_1}, x_{k+1}) \cdots k(x_{w_r+1}, x_{k+1})]^T \tag{33}
\]

The updating of the eigenvector \( V'_{k+1} \) becomes

\[
V'_{k+1} = \sum_{i=1}^{r+1} \alpha'_{k+1,i} \Phi(x_{w_i}) \tag{34}
\]

where \( w_{r+1} = k + 1 \).

According the matrix updating rule in equation 32, we have

\[
\beta_{k+1} = (K'^{-1}_{k+1})^{-1} k'(x_{k+1}) \tag{35}
\]

The parameters of the reduced KPCA model are updated and introduced to the ORKGLRT (EKG) statistic for online leak detection purpose,

\[
\text{EKG}(x_{k+1}) = \frac{1}{1 - \Phi^T(x_{k+1}) (\Phi(X'))^T P' \Gamma^{-1} P'^T (\Phi(X'))^T \Phi(x_{k+1})} \tag{36}
\]

where \( \Phi X' = D_{k+1}, \Gamma = \text{diag}(\lambda_{k+1,1}, \ldots, \lambda_{k+1,N}) \), \( P' = \begin{bmatrix} \alpha_{k+1}', \ldots, \alpha_{k+1,\ell_{r+1}}' \end{bmatrix} \) and \( \ell_{r+1} \) is the number of retained KPCS using the dictionary \( D_{k+1} \), and EW-ORKGLRT statistic (EKG) is defined as

\[
\text{EKG}(k+1) = \text{EKG}(k+1) + (1 - \varepsilon) \text{EKG}(k) \tag{37}
\]

where, \( \varepsilon \) is the smoothing parameter between 0 and 1. \( \varepsilon \) is the weight that defines the trade-off between the EKG and EKG(k) indices. The threshold for EW-ORKGLRT chart is updated as,

\[
\text{EKG}_\alpha(k+1) = g_{\text{EKG}_k} \chi^2_{h_{\text{EKG}_k}, \alpha} \tag{38}
\]

where \( g_{\text{EKG}_k} = \frac{b_{\text{EKG}_k}}{t_{\text{EKG}_k}} \) and \( h_{\text{EKG}_k} = \frac{2a_{\text{EKG}_k}}{t_{\text{EKG}_k}} \), \( b \) and \( a \) are the variance and the mean of the EW-ORKGLRT index.

### IV. LEAK DETECTION IN WATER DISTRIBUTION NETWORKS

The EW-ORKGLRT is developed in order to improve the leak detection abilities. The effectiveness of the proposed leak detection technique is assessed and compared to the ORKGLRT and KGLRT techniques in terms of three detection criteria:
1) False alarm rate (FAR) (%): percentage of wrong leak declared in leak free region,
2) Good detection rate (GDR) (%): percentage of leaky observations undetected,
3) CPU-time (CT): The time required for leak detection.

A. Water distribution system description

V. Leak detection in water distribution networks

The EW-ORKGLRT is developed in order to improve the leak detection abilities.

A. Water distribution system description

A water distribution network (WDN) is a complex process made up of hydraulic elements connected together (including reservoirs or tanks, and consumption nodes), connected by interconnecting links (composed of pipes, pumps and valves). The leak detection is evaluated using Hanoi benchmark [24]. The network diagram is shown in figure 1. The network is built using a 34 pipes and 32 nodes arranged in two branches and three loops. The process is gravity fed by a single reservoir. At this stage, the goal is to detect the leaks in the network using the developed technique. In this study, a simulation of 24 hours with a sample time of 15 minutes is considered. The network contains 31 demand nodes with index from 2 to 32 (see Figure 1). A set of sensors in the network placed in nodes 3, 10, 16, 23 and 25 (see Figure ref sensor). The KPCA model is built from these nodes. The $X$ data is obtained using the five measured pressures in nodes 3, 10, 16, 23 and 25, and is divided into training and testing data sets of 97 observations each, in order to carry out leak detection. Statistical confidence limits are set as 95%. The performance of the developed EW-ORKGLRT leak detection method is illustrated and compared to KGLRT and ORKGLRT. The comparison is assessed through different leaks. In this case, a leak is simulated in node 1 between the samples 30 and 97. The leak detection results using KGLRT, ORKGLRT and EW-ORKGLRT are presented in Figures 2, 3 and 4, respectively. If the statistic values are higher than confidence limit values then there are leaks in the process. We can show from Figure 2 that there is no leak in the system as expected when using the KGLRT chart. However, the ORKGLRT and EW-ORKGLRT methods detect clearly the leak between 30 and 97, this might be due to the adaptation of the threshold for both of them.

The leak detection results are illustrated in Table I in terms of FAR, GDR and CPU-time values with a leak in node 1. The results show that both ORKGLRT and EW-ORKGLRT provide better leak detection performance than the KGLRT and are able to detect the leak in node 4. The control chart obtained using the KPCA model is not adapted according to process changes. However, as shown in Table I, applying reduced KPCA based adaptive control chart to the same samples, pro-
vides better capabilities of adaptation to real behaviour changes of the process. These results show also that the CPU-Time needed for leak detection using ORKGLRT and EW-ORKGLRT is approximately 3300 and 2320 times faster than using KGLRT technique, respectively. This fact is useful in many industrial applications where the updating procedure requires to be processed with other steps in a short time.

<table>
<thead>
<tr>
<th></th>
<th>FAR</th>
<th>GDR(%)</th>
<th>CPU-Time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KGLRT</td>
<td>0</td>
<td>2.9412</td>
<td>0.8125</td>
</tr>
<tr>
<td>ORKGLRT</td>
<td>4.4853</td>
<td>82.3529</td>
<td>4.480610</td>
</tr>
<tr>
<td>EW-ORKGLRT</td>
<td>0</td>
<td>97.0588</td>
<td>2.453210</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

In this paper, we developed a novel technique for detecting the leaks in water distribution network. In the proposed approach, the modeling phase was addressed using reduced kernel PCA method, and the leaks were detected using statistical hypothesis test. The results demonstrated the effectiveness of the proposed framework over the conventional techniques. The detection abilities were evaluated in terms of false alarm rate, good detection rate and CPU-time.

REFERENCES