Nonparametric Kullback-divergence-PCA for intelligent mismatch detection and power quality monitoring in grid-connected rooftop PV

Azzeddine Bakdi a, *, Wahiba Bounoua b, Saad Mekhilef c, Laith M. Halabi c

a Department of Mathematics, University of Oslo, 0851, Oslo, Norway
b Signals and Systems Laboratory, Institute of Electrical and Electronics Engineering, University M’Hamed Bougara of Boumerdes, Avenue of Independence, 35000, Boumerdes, Algeria
c Power Electronics and Renewable Energy Research Laboratory (PEARL), Department of Electrical Engineering, Faculty of Engineering, University of Malaya, 50603, Kuala Lumpur, Malaysia

**A B S T R A C T**

In parallel to sustainable growth in solar fraction, continuous reductions in Photovoltaic (PV) module and installation costs fuelled a profound adoption of residential Rooftop Mounted PV (RMPV) installations already reaching grid parity. RMPVs are promoted for economic, social, and environmental factors, energy performance, reduced greenhouse effects and bill savings. RMPV modules and energy conversion units are subject to anomalies which compromise power quality and promote fire risk and safety hazards for which reliable protection is crucial. This article analyses historical data and presents a novel design that easily integrates with data storage units of RMPV systems to automatically process real-time data streams for reliable supervision. Dominant Transformed Components (TCs) are online extracted through multiblock Principal Component Analysis (PCA), most sensitive components are selected and their time-varying characteristics are recursively estimated in a moving window using smooth Kernel Density Estimation (KDE). Novel monitoring indices are developed as preventive alarms using Kullback-Leibler Divergence (KLD). This work exploits data records during 2015–2017 from thin-film, monocrystalline, and polycrystalline RMPV energy conversion systems. Fourteen test scenarios include array faults (line-to-line, line-to-ground, transient arc faults); DC-side mismatches (shadings, open circuits); grid-side anomalies (voltage sags, frequency variations); in addition to inverter anomalies and sensor faults.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Solar plants are continuously expanding as global solutions for clean energy. In the next few decades, renewable energy is going to contribute to a significant proportion of the world’s electricity needs. According to financial reports, international businesses exhibit an increasing interest in buying more renewable energy to the extent that 36 corporations, government agencies, and universities have agreed to buy 3.3 Gigawatts (GW) of wind and solar power in 2018 alongside the deals of 4.8 GW in 2017 [1]. Google, the giant company, announced in 2017 that it had met the target made in 2012 to achieve a 100% consumption of clean energy by its establishments around the world [2]. This followed the deals closed to purchase 3 GW of renewable energy capacity that year [3]. In Europe, the five largest countries in electricity production (UK, Germany, France, Italy and Spain) produced 90.5 Terawatt-hours (TWh) from solar in 2015 and had 90.4 GW of installed solar capacity at the end of 2017. Worldwide, the New Policies Scenario of the World Energy Outlook 2017 is expecting a solar Photovoltaic (PV) capacity of 2067 GW by 2040, producing 3162 TWh. And in the Sustainable Development Scenario, a goal of 3246 GW solar PV capacity is set to be achieved [4].

Many countries have already reached grid parity for solar Photovoltaics (PVs) [5], for which solar power plants will continue to get built at utility-scale, but in fact, the millions of Rooftop Mounted PV (RMPV) installations make the real potential for solar fraction [6]. RMPV installations are promoted for economic [7], social [8], and environmental [9] factors. By installing RMPV solar panels, consumers will pay less to the electric utility and may even become energy producers rather than consumers only. In addition to bill savings [10], RMPV systems contribute to lowering the demand for
fossil fuels and greenhouse gas emissions [11], they also reduce the Levelized Cost Of Electricity (LCOE) and the dependency on the utility grid [12]. The RMPV modules and energy conversion units are frequently subject to faults which if remain undetected, they cause safety hazards [13] for the personnel and fire risk [14,15]. Moreover, distribution companies are anxious about what is being injected into their grids since local malfunctions in the system cause serious problems in the grid side. Grid connection faults compromise power quality [16] and cause many grid voltage regulation issues [17], these pose several protection-related challenges [18] to avoid islanding, tripping, and interference [19].

The fast growth of sustainable energy production relies on the developments in the technology involved. Besides, the economics of PV systems play an important role in making the technology available, where an economically valuable PV system is reliable, long-lasting, and rarely prone to malfunctions. Solar plants are degrading faster than expected as a result of various defects; prompt detection of such defects in RMPV systems can guarantee the continuous healthy operation and reduce energy losses. An inclusive survey on defects in grid-connected PV systems and the most recent fault diagnosis schemes proposed in the literature is provided in Ref. [20]. Grid-connected PV systems are subjected to defects that generally occur due to equipment failures such as PV array and Maximum Power Point Tracking (MPPT) faults at the DC side, faults on the AC side (or faults at the grid level), the DC/AC inverters interface, and sensors faults as classified by Ref. [21] for which nondestructive inspection, testing, and evaluation are highly required as reported by Ref. [22]. Most of the methods proposed so far for monitoring PV systems are summarized in a review paper [23]. The authors emphasized the importance of early fault detection to prevent the risks of energy loss and disastrous fires at PV installations. Satellite-based observations are used in a remote monitoring method and compared with the expected ones to detect small grid-connected PV systems failures [24]. This method utilises solar irradiance information derived from satellites to simulate the energy yield of a PV system, which is undoubtedly less accurate than on-site measurements. Another technique [25] is based on multiple online models corresponding to different ranges of solar irradiation to predict the AC power generation that requires climate measurements; hence, this method is cost-ineffective since it involves additional sensors. The last hardware-based approaches [24,25] are known for their increased expenses of installing and maintaining such equipment and sensors which are also subject to failures and add to the complexity of the system. On the other side, the authors in Ref. [26] proposed a fault detection framework employing the fuzzy logic systems interface and artificial neural networks (ANN) techniques to detect different types of faults. However, those methods need further specifications on the faulty data. Additionally, outlier detection rules were adopted for monitoring PV systems in Refs. [27,28]; under normal operation, these rules were found to trigger false alarms and were computationally expensive. A major drawback of Artificial Intelligence (AI) methods, such as [26–28], is the requirement of labelled data in training and calibration. This, however, is practically not feasible since representative data cannot be collected during faulty operations.

In this paper, a data-driven algorithm is proposed to detect the different types of faults in grid-connected RMPV systems. Measured data of several years of operation is used for three interconnected systems, namely thin-film, monocrystalline, and polycrystalline RMPV energy conversion systems. The datasets available from the RMPV sub-systems are statistically modelled using Multiblock Principal Component Analysis (MPCA). MPCA [29,30] consists of calculating the multivariate models of each block using the standard PCA method after dividing the variables into relevant blocks. The blocking technique utilises just the correlations between the features in the particular block to estimate the scores, while the whole number of variables is used to estimate the scores in the standard PCA. Constructing the blocks typically depends on the system structure [31], considered the process sections to divide the variables into blocks that describe a unit or a specific physical or chemical operation. The measured data is projected on the MPCA model to obtain reference and online Transformed Components (TCs) which describe the amount and direction of variation in a large d-dimensional space which is orthogonal. The traditional PCA and its variants are proved effective and computationally efficient for monitoring and fault detection purposes [32,33], however, they rely on heavy assumptions such as system linearity and time-invariance (process stationarity) in the construction stage while the analysis of the principal components is based on the assumption of data following a multivariate Gaussian distribution. Unfortunately, the three assumptions do not hold in practice as it will be verified and proved experimentally in this paper.

Therefore, the most sensitive features are selected to detect and measure any operation deviation using an information gain named Kullback-Leibler (KL) divergence. Recently, KL divergence has proved its effectiveness in a considerable number of research items in multivariate process monitoring [34]. However, the KL divergence method is inefficient due to its high computational complexity that limits the application scope to low frequency (3-min sampled) data as reported in Ref. [34]. [35,36] employed univariate KL divergence on multivariate principal scores obtained from a PCA model, this procedure was extended in Ref. [37] to incipient faults. In this work, such techniques are referred to as parametrised approximations to KL divergence since they are based on assumptions of Gaussian distributions [35–37] and Gamma distributions [38] for the original data and extracted scores. The divergence in such parametrised approaches is turned into a simple detection of changes in statistical parameters such as the process mean shift and statistical dispersion which deteriorates the design sensitivity and robustness. The presented experimental analysis will prove that such approximations are inaccurate and practically far in grid-tied RMPV systems, for which a nonparametric but computationally-efficient method must be adopted. In this work, the scores are online evaluated through a sliding window approach employing KL measure through the non-parametric smooth Kernel Density Estimation (KDE) without any assumptions on the real system or its collected data. The developed approach is based on multiblock PCA decomposition and sensitive components selection followed by actual recursive KDE and accurate KL divergence. It is proved very efficient and effective since it respectively avoids the computation burden of multivariate KL divergence and escapes the basic assumptions of PCA. To this end, a wide range of tests are implemented in this article through fourteen scenarios based on real data records. A deep analysis reflects the violation of theoretical assumptions associated with traditional approaches. The obtained results prove the potential application of the proposed developments compared to state-of-the-art methods in fault detection [39,40].

The rest of this article is organised as follows: Section 2 details the scope of this work with descriptions of the different PV systems under study as well as their collected data and test scenarios; Section 3 then summarizes the design procedure of the proposed algorithms, which are then applied on the given systems in Section 4 and tested on the fourteen scenarios; the obtained results are discussed and compared; and finally, important remarks are drawn in a conclusion in Section 5.
2. Systems description and scope

Today's rooftop solar arrays do not only generate clean energy and reduce the dependency on grid power, but they must also provide a long-term sustainable and reliable power source, and they need to have a long-lasting solid foundation. On the dark side, distribution companies are anxious about what is injected in their grids and mainly the power quality. Moreover, safety hazards of the system are mapped into unsafety of the personnel working or living under such utilities. Minor drawbacks of rooftop solar systems are due to the technical standards of connectivity, accessibility, and increased maintenance costs. Because of these peculiarities, RMPV systems need to be equipped with the most reliable protection schemes and need to be continuously monitored. Despite the abundant methods available in the literature, not all of them can accurately address this problem in practice.

A rooftop mounted PV system connected to a microgrid, is considered in this article, with collected data records over the three years 2015–2017. This medium-size installation can be seen as a collection of three main interconnected subsystems as shown in Fig. 1. Their solar PV arrays are Poly-Crystalline, Mono-Crystalline, and thin film, they respectively consist of 16, 27, and 20 units as shown in the rooftop installations is Fig. 2(a), their rated power is 2 kW, 2.025 kW, and 2.7 kW, respectively. In this article, \( S_1 \), \( S_2 \), and \( S_3 \) refer to the three subsystems in their respective order. The three blocks are connected to a microgrid, powering various loads of the research laboratory and synchronized with local sources and with the main grid lines though two SUNNY BOY 1600 TL inverters, and SUNNY BOY 2500HF inverter as shown in Fig. 2(b). Technical specifications of the SUNNY BOY 1600 TL inverter can be found in its operating manual [41] and its ratings are listed in the technical datasheet [42]. Technical specifications and ratings of the SUNNY BOY 2500HF inverter are also provided in Refs. [43,44].

The three subsystems are also connected to SMA SUNNY SENSORBOX [45,46] to measure environmental conditions such as solar irradiance, wind speed, ambient temperature, and module temperature. The sensor box is powered through SMA power injector and integrated through a communication bus with SMA SUNNY WEBBOX [47,48] which records data from all connected devices (sensor box and grid-tied inverters). The latter is connected to a local network and desktops to store and monitor data measurements.

Recorded data include measurements of over 60 system variables (referred to as “measured values” in the technical description [49,50]), operating parameters, log events, and messages. These are listed and described in Refs. [49,50] for both inverter types. In the developed monitoring algorithm, monitoring variables are limited to the 27 fault-relevant signals as summarized in Table 1. Variables one to four are the environmental conditions measured through the sensor box, this set of variables is common for the three blocks. In addition to the external measurements, subsystems \( S_1 \), \( S_2 \), and \( S_3 \) are respectively monitored by variables 5 to 11, 12 to 18, and 19 to 27.

Analysis of a PV system performance is highly dependent on the environmental conditions, and mainly the actual solar irradiance as well as ambient and module temperature. The influence of variations in temperature and irradiance on the PV module parameters are discussed in Ref. [51] and thermal performance of PV modules is discussed in Ref. [52]. Authors in Ref. [15] demonstrated the strong correlation between an increase in the temperature and the change in energy production by rooftop integrated PV panels. In practice, a key aspect of environmental signals is their large (natural) variability, measurement errors, and noise as demonstrated in Fig. 3 which represents the actual measurements recorded over one day. The DC-side signals, as shown in Fig. 4, are highly and nonlinearly correlated with the latter variables and consequently exhibit large variability as well. In addition to the large variability within electrical signals over a day, the energy produced by the different PV arrays also exhibits a non-negligible variation over a year as shown in Fig. 5 for 2017.

Such strong correlations, large variations, and uncertainties cannot be ignored in practice when designing a monitoring device. Moreover, these conditions play a major role in masking symptoms of possible mismatches in the system operation, small anomalies,
Fig. 2. Main components of the rooftop-mounted PV system. (a) Photovoltaic modules, (b) Grid-tied solar inverters.

Table 1
Description of the selected monitoring variables of the PV system.

<table>
<thead>
<tr>
<th>Sensor Box</th>
<th>Name</th>
<th>Detailed description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IntSolIrr</td>
<td>Total irradiation on sensor surface</td>
<td>W/m²</td>
</tr>
<tr>
<td>2</td>
<td>TmpAmbC</td>
<td>Environment (Ambient) temperature</td>
<td>°C</td>
</tr>
<tr>
<td>3</td>
<td>TmpMdulC</td>
<td>PV module temperature</td>
<td>°C</td>
</tr>
<tr>
<td>4</td>
<td>WindVelms</td>
<td>Wind speed</td>
<td>m/s</td>
</tr>
<tr>
<td>Subsystem 1</td>
<td>Fac</td>
<td>Power frequency</td>
<td>Hz</td>
</tr>
<tr>
<td>5</td>
<td>IacIst</td>
<td>Grid current</td>
<td>mA</td>
</tr>
<tr>
<td>6</td>
<td>Ipv</td>
<td>DC current input</td>
<td>mA</td>
</tr>
<tr>
<td>7</td>
<td>Pac</td>
<td>AC active power across all phases</td>
<td>W</td>
</tr>
<tr>
<td>8</td>
<td>Uac</td>
<td>AC voltages (average of all string voltages)</td>
<td>V</td>
</tr>
<tr>
<td>9</td>
<td>Upv-Ist</td>
<td>DC voltage input</td>
<td>V</td>
</tr>
<tr>
<td>10</td>
<td>UpvSolll</td>
<td>Reference voltage</td>
<td>V</td>
</tr>
<tr>
<td>Subsystem 2</td>
<td>Fac1</td>
<td>Power frequency</td>
<td>Hz</td>
</tr>
<tr>
<td>12</td>
<td>IacIst1</td>
<td>Grid current</td>
<td>mA</td>
</tr>
<tr>
<td>13</td>
<td>Ipv1</td>
<td>DC current input</td>
<td>mA</td>
</tr>
<tr>
<td>14</td>
<td>Pac1</td>
<td>AC active power across all phases</td>
<td>W</td>
</tr>
<tr>
<td>15</td>
<td>Uac1</td>
<td>AC voltages (average of all string voltages)</td>
<td>V</td>
</tr>
<tr>
<td>16</td>
<td>UpvIst1</td>
<td>DC voltage input</td>
<td>V</td>
</tr>
<tr>
<td>17</td>
<td>UpvSoll1</td>
<td>Reference voltage</td>
<td>V</td>
</tr>
<tr>
<td>Subsystem 3</td>
<td>AMsAmp</td>
<td>DC current input in A</td>
<td>A</td>
</tr>
<tr>
<td>19</td>
<td>AMsVol</td>
<td>DC voltage input in V</td>
<td>V</td>
</tr>
<tr>
<td>20</td>
<td>AMsWatt</td>
<td>DC power input in W</td>
<td>W</td>
</tr>
<tr>
<td>21</td>
<td>GridMsAphsA</td>
<td>Grid current phase L1 in A</td>
<td>A</td>
</tr>
<tr>
<td>22</td>
<td>GridMsHz</td>
<td>Grid frequency in Hz</td>
<td>A</td>
</tr>
<tr>
<td>23</td>
<td>GridMsPhVphsA</td>
<td>Grid voltage phase L1 in V</td>
<td>A</td>
</tr>
<tr>
<td>24</td>
<td>GridMsTotVA</td>
<td>Total apparent power in VA</td>
<td>VA</td>
</tr>
<tr>
<td>25</td>
<td>GridMsWphsA</td>
<td>Active power phase L1 in W</td>
<td>W</td>
</tr>
<tr>
<td>26</td>
<td>Pac2</td>
<td>Delivered active power in W (total)</td>
<td>W</td>
</tr>
</tbody>
</table>

Fig. 3. One-day variation of environmental conditions.
or unmatched power quality. These observations receive particular
attention in section 3 when modelling the primary correlations of
those variable with internal and grid signals through a nonpara-
metric design. Comparisons with conventional schemes, ignoring
some of such major factors, are drawn in section 4 to demonstrate
the significance of such recommendations.

In this article, data records of the described system span three
years of continuous operation during 2015—2017. Various scenarios
are investigated in this article, as summarized in Table 2. These
scenarios span physical as well as electrical and environmental
faults and mismatches at the PV arrays, inverters, and grid levels.
Due to the prementioned factors, the DC-side faults are hard to be
detected however they have a medium severity level because of the
bounded voltage and current of the PV modules. On the contrary,

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>Fault</th>
<th>location</th>
<th>Behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Array faults (DC- side faults)</td>
<td>F1</td>
<td>Ground fault</td>
<td>$S_1$</td>
<td>Abrupt</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>Line to line fault</td>
<td>$S_2$</td>
<td>Abrupt</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td>Parallel arc faults</td>
<td>$S_1$</td>
<td>Abrupt</td>
</tr>
<tr>
<td></td>
<td>F4</td>
<td>Series arc faults</td>
<td>$S_1$</td>
<td>Abrupt</td>
</tr>
<tr>
<td></td>
<td>F5</td>
<td>Partial shading faults</td>
<td>$S_2$</td>
<td>Abrupt</td>
</tr>
<tr>
<td></td>
<td>F6</td>
<td>Short circuit</td>
<td>$S_3$</td>
<td>Abrupt</td>
</tr>
<tr>
<td></td>
<td>F7</td>
<td>Open circuit</td>
<td>$S_1$</td>
<td>Abrupt</td>
</tr>
<tr>
<td>Inverter faults</td>
<td>F8</td>
<td>Frequency control (overshoot)</td>
<td>$S_1$</td>
<td>Incipient</td>
</tr>
<tr>
<td></td>
<td>F9</td>
<td>Slow frequency control (transient)</td>
<td>$S_1$</td>
<td>Incipient</td>
</tr>
<tr>
<td>Grid faults (AC- side faults)</td>
<td>F10</td>
<td>Voltage sag</td>
<td>$S_1$, $S_2$, $S_1$</td>
<td>Intermittent</td>
</tr>
<tr>
<td></td>
<td>F11</td>
<td>Frequency variation</td>
<td>$S_{1,2,3}$</td>
<td>Random</td>
</tr>
<tr>
<td>Sensor faults</td>
<td>F12</td>
<td>Voltage sensor fault</td>
<td>$S_1$</td>
<td>Random</td>
</tr>
<tr>
<td></td>
<td>F13</td>
<td>Temperature sensor fault</td>
<td>$S_1$</td>
<td>Bias</td>
</tr>
<tr>
<td></td>
<td>F14</td>
<td>Current sensor fault</td>
<td>$S_1$</td>
<td>Random</td>
</tr>
</tbody>
</table>
AC-side faults are less challenging to detect, however being highly severe, they need to be detected at their very early stages to ensure a safe microgrid integration and match the technical requirements and desired power quality [53]. provides a comprehensive review of a wide range of faults in grid-connected PV systems, causes, symptoms, and their respective protection schemes [54]. highlighted the power quality issues for building integrated renewables such as PV systems. Line to line, line to ground, and short circuit faults, in general, occur between two points of different potentials. These faults occur due to breakdown of insulation, corrosion of inductors, maintenance errors, and damage inside the PV arrays. Arc faults [55] are short-term versions of those faults and may cause permanent faults. These faults contribute to serious fire threats and safety hazards [53]. Open circuit faults may occur after some of the previous faults, and together with shading faults are considered as mismatches. Those scenarios are discussed in section 4, showing a correlation between the detection delay and the severity of each situation.

3. Data analysis and monitoring algorithms

In long terms, digitized PV plants generate vast amounts of data which reflects the historical behaviour and sharpest details within the system. Such valuable information can be exploited for artificial modelling of the PV system [56], parameter estimation and monitoring [57], and even PV power generation forecasting [58]. A common fact is that such high-dimension data exhibits a lower statistical rank and it varies in a lower-dimensional space. Moreover, the PV system data is multivariate and includes weak as well as strong parallel and serial correlations among its variables; the data also exhibits large (DC-side) and small (grid-side) variations which are all of a significant importance and in which various anomalies exhibit different patterns to be detected by a protection system. Such information can be extracted using PCA while decorrelating the initial space to reduce the dimension of the treated problem and therefore its computational cost while considering a univariate analysis of the most sensitive TCs. Multiblock decomposition [59,60] is used in this work for factorization and decentralized monitoring. Instead of statistical preferences, the system variables are divided into 3 blocks according to the given structure of the rooftop mounted solar system, as demonstrated in the previous section. Each subsystem is represented by a block $S_b$ for $b = 1, 2, 3$, several datasets are acquired for analysis purposes, one healthy set is recorded for constructing the basic model and it is then validated and tested on a set of 14 scenarios of design and quality mismatches that the system is subject to. The healthy dataset is measured during fault-free operation, down to zero mean ($\mu$) and unit variance ($\sigma^2$), and hence all the variables can be treated equally during the analysis [61]. The auto-scaled data matrix is obtained as:

$$X_{S_b} = \begin{bmatrix} x_{b,1} - \mu_{b,1} / \sigma_{b,1} & x_{b,2} - \mu_{b,2} / \sigma_{b,2} & \cdots & x_{b,N_b} - \mu_{b,N_b} / \sigma_{b,N_b} \end{bmatrix}$$

(1)

where $\mu_i$ and $\sigma_i$ are the mean and the standard deviation of the $i^{th}$ variable of the healthy data set in subsystem $S_b$, these can be directly estimated or updated at any stage as follows:

$$\mu_{b,i} = \frac{1}{N_b} \sum_{i=1}^{N_b} x_{b,i} \quad \text{for} \quad b = 1, 2, 3, \quad i = 1, \ldots, l_b$$

$$\sigma_{b,i}^2 = \frac{1}{N_b} \sum_{i=1}^{N_b} (x_{b,i} - \mu_{b,i})^2 \quad \text{for} \quad b = 1, 2, 3, \quad i = 1, \ldots, l_b$$

(2)

(3)

For simplicity, the normalized data matrix $X_{S_b}$ is denoted as $X_{b}$ and is given by:

$$X_{b} = [x_{b,1} x_{b,2} \cdots x_{b,l_b}] \in \mathbb{R}^{N_b \times l_b}$$

(4)

The data transformation is based on the sample covariance matrix $\Phi_b$ of the data, where:

$$\Phi_b = \frac{1}{N_b - 1} X_{b}^{T} X_{b}$$

(5)

through the spectral decomposition of the later as:

$$\Phi_b = P_b A_b P_b^T$$

(6)

henceforth, the block loading matrix $P_b$ is obtained with orthogonal components, i.e., $P_b P_b^T = I_{l_b}$, constructed by the eigenvectors which represent the variations directions. $A_b = \text{diag}(\lambda_{b,1}, \lambda_{b,2}, \ldots, \lambda_{b,l_b})$ is a diagonal matrix constructed of the eigenvalues in a decreasing order ($\lambda_{b,1} \geq \lambda_{b,2} \geq \ldots \geq \lambda_{b,l_b} \geq 0$), each eigenvalue represents the amount of variance per the corresponding direction.

Subsequently, the data collected from the PV system blocks is transformed by PCA projection into a new matrix $T_{A_b} \in \mathbb{R}^{N_b \times l_b}$ named as block score matrix of uncorrelated variables $(t_{ab,1}, t_{ab,2}, \ldots, t_{ab,l_b})$:

$$T_{A_b} = X_{b} P_{b}$$

(7)

PCA allows the provision of a set of uncorrelated variables from the original set of correlated variables. These are called Transformed Components (TCs). At this stage, $l_b$ reference TCs components are obtained for each subsystem $S_b$ for $b = 1, 2, 3$. PCA results in a statistical model that describes the PV system data patterns and correlations given the reference data profile. Moreover, the resulting TCs (combinations of the correlated data) are independent and can be monitored individually in real-time.

3.2. Smooth KDE & Kullback-Leibler divergence

Considering the high-level uncertainty and the large variability in the PV system data compared to symptoms of anomalies as mentioned in the previous section, the Kullback-Leibler divergence (KL-divergence) [62] is adopted in this article. The idea is to develop robust and sensitive measures of any deviation in the overall PV system performance at time instance $n$ from the nominal operation described explicitly by historical data. For feasible computation time and resources, and for accurate monitoring purposes, this measure is calculated in the decorrelated space and selecting only the most sensitive components. The KL-divergence is an information-based measure of dissimilarity between two probability distributions $f_X(x)$ and $\tilde{f}_X(x)$ defined over the same random variable $X$. KL-divergence is a special case of $\alpha$-divergence functions and it is asymmetrical non-negative quantity i.e. $\text{DKL}[f_X(x) : \tilde{f}_X(x)] = \text{DKL}[\tilde{f}_X(x) : f_X(x)] \geq 0$ [63].
The KL-divergence between two probability density distributions \( f_X(x) \) and \( f_X(x) \) is defined as the expectation over \( f_X \), and it is given by:

\[
D_{KL}(f_X(x) : f_X(x)) = E_{f_X}[\log \frac{f_X(x)}{f_X(x)}].
\] (8)

If \( X \) is a discrete random variable then Eq. (8) reduces to:

\[
D_{KL}(f_X(x) : f_X(x)) = \sum_{x \in X} f_X(x) \log \frac{f_X(x)}{f_X(x)}
\] (9)

And for continuous Random Variable \( X \):

\[
D_{KL}(f_X(x) : f_X(x)) = \int f_X(x) \log \frac{f_X(x)}{f_X(x)} dx
\] (10)

It is a measure of the inefficiency of assuming that the distribution is \( f_X \) when the true distribution is \( f_X \). In this work, \( f_X \) represents the reference density function created in an offline stage through the reference TCs, while \( f_X \) is the online-estimated test density function. The idea is used in this article to measure the divergence of a prevailing PV system behaviour and characteristics according to its most recent measurements for a reference data profile.

KL-divergence is widely proved efficient for monitoring purposes, authors in Ref. [37] have obtained a closed-form approximation for this measure across variable following Gaussian distribution [64]. Unfortunately, the highly sensitive information gain, in this situation, is turned into measuring deviation in the mean and variance only, this heavy assumption deteriorates the performance of this measure since data in practice does not follow a Gaussian distribution, particularly during the occurrence of anomalies. Further assumptions are made in the prementioned approach such as the limitations to linear static (time-invariant) systems. A second drawback is the density ratio estimation involves multiple parameters with optimization functions, this approach is computationally infeasible for large-scale systems, especially if the system has fast dynamics. Alternatively, another approximation of such measure is widely used in the literature based on direct density ratio estimation, called the importance estimation, used in Refs. [65,66]. This approach reduces relatively the computation cost with light assumptions however the estimated ratio is still multivariate and demanding, furthermore, the ratio could explode to infinity.

The prementioned lacks promote motivations of this article to design a novel approach to measure the KL-divergence to preserve its high sensitivity which is highly crucial for identifying early signs of anomalies during power generation. The multivariate problem is turned into a univariate analysis which greatly reduces the computation time and resources, moreover, monitoring only sensitive TCs greatly improves the algorithm performance.

Without any assumptions on the system or its data, the Probability Density Functions (PDFs) are recursively estimated for the block TCs by a means of a non-parametric estimation method called the Kernel Density Estimation (KDE), also known as Parzen windows [67] that provides a smooth estimate based on sufficient amount of data. Given a set of offline-estimated reference block TCs, represented as follows:

\[
T_{k,n} = \{t_{b1,n}, t_{b2,n},\ldots, t_{bN_b,n}\}, \quad t_{b1,n} \in \mathcal{R}^{N_b \times 1}
\] (11)

for all subsystems, where \( k \) is the time index. Hence \( t_{b,n,k} \) is the \( k \)th reference TC of sub-block \( S_b \) at time \( k \), and the reference PDFs \( f_{t_{b,n,k}}(t; h) \) can be estimated through KDE as:

\[
\hat{f}_{t_{b,n,k}}(t; h) = \frac{1}{N_h} \sum_{i=1}^{N_h} K\left(\frac{t - t_{b,n,k}^i}{h}\right)
\] (12)

for \( i = 1, \ldots, N_h \) for the three subsystems \( b = 1, \ldots, 3 \) using the smoothing kernel function:

\[
K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)
\] (13)

\( h > 0 \) is the bandwidth that controls the smoothing and the estimation. This work adopts a bandwidth that minimizes the Mean Integrated Squared Error (MISE) [68]:

\[
\text{MISE}(h) = \mathbb{E} \left[ \int \left( \hat{f}_{t_{b,n,k}}(t; h) - f_{t_{b,n,k}}(t; h) \right)^2 dt \right]
\] (14)

3.3. Performance monitoring indices

At any time index \( n \), the measured data is first scaled using Eq. (1), however, using the same parameters obtained in Eq. (2) and Eq. (3). The scaled measurement is projected on the multiblock PCA models to estimate the online block TCs. The test TCs at a time index \( n \), are formed by augmenting the \( N_e \) most recent samples of the TCs.

\[
T_{k,n} = \{t_{b1,n}, t_{b2,n},\ldots, t_{bN_b,n}\}, \quad t_{b1,n} \in \mathcal{R}^{N_b \times 1}
\] (15)

where \( n_0 = n - N_e + 1 \), for the three system blocks \( b = 1, 2, 3 \).

The test block TCs are observed within a sliding window of length \( N_e \), these projections (Eq. (7)) are stacked with their projections (Eq. (14)) as a monitoring statistic. The normal state, defined by the null hypothesis \( H_0 \), is characterized by a region of non-significance where the real-time divergence is within a pre-established threshold \( \delta \). Whenever\( \delta \) diverges significantly from the region of acceptance, the owner of the RMPV system or a supervision platform would be inclined to reject the null hypothesis and accept the alternative hypothesis \( H_A \). The latter situation indicates the departure from nominal to the abnormal state of the PV system. Consequently, the following decision rule is formulated:
\[ \text{DKL} \left[ \hat{f}_{\text{AB}}(t; h); \hat{f}_{\text{B}}(t; h) \right] \overset{\mathcal{H}_a}{\geq} \delta \]

(17)

Through this statement, a decision is made over the KL-divergence of the score components of each sub-block as follows:

\[ \begin{aligned}
&\mathcal{H}_0: \text{DKL} \left[ \hat{f}_{\text{AB}}(t; h); \hat{f}_{\text{B}}(t; h) \right] \leq \delta_{b_i} \\
&\mathcal{H}_a: \text{DKL} \left[ \hat{f}_{\text{AB}}(t; h); \hat{f}_{\text{B}}(t; h) \right] > \delta_{b_i} 
\end{aligned} \]

(18)

for \( b = 1, 2, 3 \) and \( i = 1, \ldots, l_b \), with respective control limits \( CL_{b_i} = \delta_{b_i} \). The previous hypothesis of Eqs. (17) and (18) will be adapted at time \( n \) to:

\[ \begin{aligned}
&\mathcal{H}_0(n) : D_{b_i}(n) \leq CL_{b_i} \text{ for } b = 1, 2, 3 \text{ and } i = 1, \ldots, l_b \\
&\mathcal{H}_a(n) : D_{b_i}(n) > CL_{b_i} \text{ for } b = 1, 2, 3 \text{ and } i = 1, \ldots, l_b
\end{aligned} \]

(19)

Notice that the null hypothesis is violated whenever any of the online block components has its test density function \( \hat{f}_{\text{B}}(t; h) \) diverged considerably (above the control limit) from its respective reference function \( \hat{f}_{\text{B}}(t; h) \). Moreover, it will be shown in the next section that monitoring the entire RMPV system is reduced to monitoring the first and last TCs which are sensitive to large variability (in DC-side and mismatches) and small variations (AC-side and quality deviation), respectively.

The sensitive indices are observed during nominal operation conditions and their control limits are constructed through nonparametric confidence intervals based on their estimated cumulative distribution functions. Allowing a tolerable level of false alarms (\( \alpha = 1\% \)), the control limits are selected to ensure a coverage probability of \( (1-\alpha)\% \) of the measured samples during healthy operation conditions are flagged as safe. The thresholds of such indices are established empirically such that the PV system is considered to be operating in normal mode if KL-divergence between the estimates of the reference density function and the online test density function is approximately zero i.e. the two distributions are similar. While any dissimilarity between the distributions will appear as a departure of DKL from the threshold and this will be regarded as an abnormality.

4. Experimental results and discussion

This section is based on the real records measured during the three years 2015–2017 from the three interconnected RMPV systems, installed on the roof of Power Electronics and Renewable Energy Research Laboratory (PEARL) of MALAYA University. The three blocks of the system, as described in Section 2, are connected to a microgrid, powering various loads of the research laboratory and synchronized with local sources and with the main grid lines depicted in Figs. 1 and 2 above. Huge datasets are available where 27 fault relevant signals are selected as detailed in Table 1, this covers all the modules, inverters, and grid measurements for the three systems. The rich datasets are filtered to remove records with erroneous and missing values and pre-processed for noise reduction, the available data is then exploited to analyse the system behaviour and used for the design and implementation and validation of the proposed algorithms as well as tests and comparisons.

Single and multiple events are investigated in 14 scenarios as described in Table 2; these are injected in different locations with various behaviours.

The data is first randomly divided into regions, the training regions for the global as well as the multiblock PCA models are of 500 samples for each signal \( X_n \in \mathbb{R}^{200 \times 27} \), while a 40000 long set is used for validation \( X_n \in \mathbb{R}^{40000 \times 27} \), while each of the 14 test scenarios includes 5000 samples \( X_n \in \mathbb{R}^{5000 \times 27} \). Multiblock PCA models are first constructed for the RMPV systems, each block data \( X_{ab}^{500} \) is first autoscaled as given in Eqs. (1)–(3), and the multiblock PCA decomposition is achieved through Eqs. (4)–(6) for the three blocks \( b = 1, 2, 3 \). The resulting TCs have their reference densities estimated through Eq. (12) with all samples of projected training data \( N_p = 500 \), the respective online TCs have their PDFs estimated at each time instance through KDE as in Eq. (16) with \( N_p = 300 \) samples, the online TCs are compared to their reference ones and evaluated based on the dissimilarity between their PDFs through KL divergence through Eqs. (10) and (17). The traditional KLD approach is expelled from performance comparison since it is completely impractical to consider recursive estimation of online to reference density ratio for such 27-dimensional space. The approximation multivariate KLD approach [34] is of high complexity for this RMPV system and data with a computation time measured around weeks in addition to out-of-memory problems.

Fig. 6 demonstrates some of the online block Transformed Components (TCs), obtained through projecting the online measurements from the RMPV systems on the reference artificial models, during normal and faulty operations of different scenarios. Notice the poor capability of the block TCs in discriminating the faulty (F) and fault-free (H) situations, more importantly, the components of the residual subspace are even more sensitive than those of the principal subspace showing some sort of separation of fault clusters. While the principal subspace describes the dominant variation due to natural behaviour, the residual subspace contains the negligible variation, which in most of the cases, is considered as attributed to noise and anomalies. The evaluation of the TCs in both subspaces must be very accurate for robust monitoring. MPCA decorrelates multivariate data from a high-dimensional space into univariate one-dimensional TCs which still describe the original covariance (variations and correlations). This is extremely advantageous but unfortunately cannot detect faults and quality deviations. The limitation of conventional PCA [32,39] and parametrised KL approaches [35–38] is investigated in the following, while nonparametric KL divergence of the same TCs in Fig. 6 will be proved very effective in the detection.

A comparison of conventional PCA statistics such as Q and \( T^2 \) [32,39] is made in Figs. 7 and 8 across scenarios 2, 7, and 8 which respectively stand for line-to-line fault, open circuit fault, and...
current sensor fault in subsystems 2, 1, and 3, respectively. Fig. 7 demonstrates the poor performance of both PCA $Q$ and $T^2$ statistics \cite{32,39} to show any symptoms before and after faults are introduced at the 1500th sample, even though these statistics combine all the TCs. Fig. 8 shows how far are the actual densities from Gaussian or Gamma distributions, it also highlights the time-varying characteristics of the PDFs across principle and residual block TCs. The Figure shows the real data projection histograms ($H$) during fault-free operation, and some smooth Kernel Density Estimates (KDE), as given in Eq. (11)–(14), at a few instances. Those KDEs are given for the indicated TCs including their PDFs during Healthy mode ($P_H$) that was used for training, these are the reference densities for those TCs of their corresponding subsystems, notice their overlapping with $H$ due to the accuracy of smooth KDE. The PDF KDEs are also given in this Figure at other independent instances: Before the Fault ($P_{BF}$), During Fault ($P_{DF}$), and After Fault ($P_{AF}$) for the three situations. Notice first that the distributions are not normal, a condition that violates a heavy assumption of PCA and its statistics, notice also that these faults are characterized by distortions in the obtained PDFs rather than deviations of the squared distance in one direction. Approximating the PDFs of TC1, TC9, and TC11 with a Gaussian distribution, the parametrised KL approaches of \cite{35,37} failed to detect fault F2 through TC1 since there is no clear mean shift from and no change in

**Fig. 6.** Normal and faulty data visualization in (a) last TCs of S1 and (b) first TCs of S2.

**Fig. 7.** The poor detection performance of conventional PCA through $Q$ and $T^2$ charts of \cite{32,39}.

**Fig. 8.** Reference and online densities of TCs before (BF), during (DF), and after (AF) faults.
A. Bakdi et al. / Energy 189 (2019) 116366

10

variance during various experiments before and after the fault. TC9 produces a false alarm due to the natural variation of density before the fault (PWy). The same results are obtained if the parametrised Gamma distribution-based KL approach [38] is applied.

Recall Fig. 6(a) where F7 cannot be classified from H along TC9, and Fig. 6(b) where TC1 cannot discriminate F2 from H. Fig. 8 shows that the dissimilarity between the PDFs of the online block TCs are estimated in the same manner as their reference densities, however, at each time new variable measurements are recorded and the dissimilarity between the reference estimated PDFs of the reference TCs and their online block counterparts is evaluated each time through Eqs. (10) and (19). The monitoring performance of the developed method is tested against all the scenarios as shown in Figs. 9–11 for RMPV subsystems S1, S2, and S3 respectively, including those shown in Figs. 7 and 8, where any performance deviation across the RMPV systems is measured with high accuracy.

Figs. 9–11 show the high sensitivity of the proposed deviation measures in addition to high robustness that can be tuned with some Control Limits of the Divergence (CLD). The fault detection alarms are then designed for the RMPV system based on the hypothesis tests as given in Eqs. (17) and (18) according to their online constructed thresholds tuned though an independent data set (Eq. (20)). The designed control limits allow for some negligible divergence attributed to measurement noise and training inaccuracies so that only a considerable divergence which is out of control at a given time point will trigger an alarm of a near-hazardous situation. The event Detection Time Delay (DTD) is then calculated as the faulty operation time before the fault is truly reported, this performance index is reported in Table 3 in addition to which monitoring index (Di) had first triggered a fault alarm. These results confirm the conclusions in Refs. [32,39] where first TCs are most sensitive to incipient and large variability faults whereas last TCs are sensitive to small intermittent and abrupt faults. Since the ith index (Di) has firstly detected the jth fault, the ith block TC is most sensitive to that fault and it is worth monitoring through nonparametric KL divergence for proper detection of that fault in the future.

While the conventional methods [32,39] and [35 to 38] failed to detect dangerous faults F2 (line to line fault), F7 (open circuit fault), and F14 (feedback current sensor fault), as shown in Figs. 7 and 8, these faults are successfully detected by the proposed method as demonstrated in Figs. 9–11. In addition to this superior detection performance, the proposed Kullback-divergence-MPCA approach ensures other potential applications over state-of-the-art methods as summarized in Table 4.

Table 4 compares the proposed and different methods according to eight practical criteria. Assumption-free methods do not rely on any analytical models or theoretical assumptions on the system and its collected data which are reflected if an approach is investigated through real applications. Another criterion is if the approach is investigated for long term reliability and the ability to handle a large amount of data to ensure satisfactory results independently from the training instance (refer to monthly variations in Fig. 5). A major concern is the requirement of labelled data for training on different types of mismatches and power quality deviations, this is an aspect of supervised approaches while the “not supervised” class includes semi-supervised and unsupervised approaches only. Due to the fast dynamics and high-frequency high-dimensional data of grid-integrated RMPV systems, computation efficiency is important to realize the online real-time mismatch detection and power quality monitoring. A KL method such as [34] requires a computation time of several weeks in such applications without problems and it cannot be used in reality even if it is theoretically proved effective. More importantly, any approach must be checked for its proved reliability of detecting different types of RMPV system faults such as array mismatches, inverter faults, and grid perturbations. The different approaches are compared according to these criteria for which, $\ast$, $\mathbf{1}$, and $\mathbf{2}$ stands for positive, negative, and not-given evaluations, respectively. Notice that the presented approach outperforms the state-of-the-art methods and exhibits the most potential applications.

![Fig. 9. Response of sensitive indices $D_i(t)$ to faults in subsystem S1](image)
5. Conclusion

This article considers the safe operation of rooftop-mounted PV installations to avoid hazardous events and ensure a smooth injection to microgrid with good power quality. The work exploits several years of real measurements of three interconnected RMPV systems: Poly-Crystalline, Mono-Crystalline, and thin-film modules with their energy conversion systems with respective capacities of 2 kW, 2.025 kW, and 2.7 kW. Performance deviation is investigated through fourteen test scenarios which span array faults such as lineFig. 10. Response of sensitive indices $D_1(t)$ to faults in subsystem $S_2$

Fig. 11. Response of sensitive indices $D_1(t)$ to faults in subsystem $S_3$

Table 3
Detection time delay (samples) across the RMPV system scenarios.

<table>
<thead>
<tr>
<th>Subsystem 1</th>
<th>Subsystem 2</th>
<th>Subsystem 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fault</td>
<td>DTD</td>
<td>$D_i$</td>
</tr>
<tr>
<td>F1</td>
<td>31</td>
<td>9</td>
</tr>
<tr>
<td>F4</td>
<td>112</td>
<td>9</td>
</tr>
<tr>
<td>F7</td>
<td>146</td>
<td>9</td>
</tr>
<tr>
<td>F10</td>
<td>47</td>
<td>9</td>
</tr>
<tr>
<td>F11</td>
<td>66</td>
<td>8</td>
</tr>
<tr>
<td>F12</td>
<td>75</td>
<td>9</td>
</tr>
</tbody>
</table>
to line, ground, and transient arc faults; DC-side mismatches in form of shadings and open circuits; grid-side anomalies such as voltage sags and frequency variations; in addition to inverter anomalies and sensor faults.

For this purpose, novel data-driven methodologies are developed for long-term performance monitoring and deviation measurement. Multiblock PCA is used in this article for statistical modelling and multivariate data decomposition and decorrelation to project the online measurements into block transformed components which are more sensitive and computationally efficient to analyse. Novel significant extensions are also proposed for accurate evaluation and robust alarm generation using Kullback-Leibler divergence through smooth kernel density estimation in a moving-window approach.

The designed algorithms are explicitly based on multivariate analysis and information gain measure, these were applied to the analysis of large datasets of real measurements and tested against the fourteen different scenarios in the RMPV systems. While theoretical methods completely failed to detect line to line, open circuit, and feedback current sensor faults in RMPV system, the presented design was proved highly effective for its assumption-free approach which successfully detected all faults in the experimented scenarios with acceptable performance. At the same time, the computationally-efficient algorithm is easily realized for online applications in RMPV systems. Moreover, the obtained results demonstrate the potential applications of the proposed strategies and outperform their conventional counterparts in terms of reliable indication of performance deviation with increased robustness and sensitivity.

<table>
<thead>
<tr>
<th>Proposed</th>
<th>[32]</th>
<th>[39]</th>
<th>[34]</th>
<th>[33]</th>
<th>[40]</th>
<th>[26]</th>
<th>[27]</th>
<th>[28]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assumption-free</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real application</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long term reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Array mismatches</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inverter faults</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grid perturbations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**References**


