Massive Streaming PMU Data Modeling and Analytics in Smart Grid State Evaluation Based on Multiple High-Dimensional Covariance Tests

Lei Chu, Robert Qiu, Fellow, IEEE, Xing He, Zenan Ling and Yadong Liu

Abstract—The analogous deployment of phase measurement units (PMUs), the increase of data quantum and the deregulation of energy market, all call for the robust state evaluation in large scale power systems. Implementing model based estimators is impractical because of the complexity scale of solving the high dimension power flow equations. In this paper, we first represent massive streaming PMU data as big random matrix flow. By exploiting the variations in the covariance matrix of the massive streaming PMU data, a novel power state evaluation algorithm is then developed based on the multiple high dimensional covariance matrix tests. The proposed test statistic is flexible and nonparametric, which assumes no specific parameter distribution or dimension structure for the PMU data. Besides, it can jointly reveal the relative magnitude, duration and location of a system event. For the sake of practical application, we reduce the computation of the proposed test statistic from \(O(cn^2)\) to \(O(nm^2)\) by principal component calculation and redundant computation elimination. The novel algorithm is numerically evaluated utilizing the IEEE 30-, 118-bus system, a Polish 2383-bus system, and a real 34-PMU system. The case studies illustrate and verify the superiority of proposed state evaluation indicator.

Index Terms—State Evaluation, Massive Streaming PMU Data, Multiple High-dimension Covariance Matrix Tests, Efficient Calculation, Smart Grid.

I. INTRODUCTION

RELIABLE operation and intelligent management of electric power systems have a heavy influence on daily life. Recently, power companies, scholars and researchers keep an eye on utilizing PMUs to improve wide area monitoring, protection, and control (WAMPAC) [1, 2]. Some large-scale implementations of synchrophasor technology for managing the power grids across the world have been brought online. As an illustration, there were about 2400 PMUs deployed in power grids in China as of 2013 [3]; North America and India have coverage from about 2000 and 1800 PMUs by 2015, respectively [4]. Accordingly, designing, monitoring, and controlling such systems are becoming increasingly more challenging as a consequence of the steady growth of their size, complexity, level of uncertainty, unpredictable behavior, and interactions [5-7].

Efforts are in place to take synchrophasor technology to evaluate power states and develop reliable operational procedures to better understand and manage the power grids with wide-area visualization tools using PMU data. These power state evaluation methods can be generally organized into two categories: model-based estimators and data driven estimators. Model-based analysis is a kind of traditional method for offline analysis of state evaluation in power systems. Lof and Anderson presented statistical state evaluation indices based on the largest singular value of the inverse of the power flow Jacobian matrix [8]. GhioceI and Chow extended the result in [8] and identified power flow control infeasibilities in a large-scale power system [9]. Pordanjani, Wang and Xu assessed the state evaluation using Channel components transform [10]. More recently, equivalent nodal analysis for power state evaluation was shown in [11]. With the help of eigenvalues, eigenvectors, and participation factors of the power flow Jacobian matrix, the system characteristics can be predicted by these estimators. However, they hardly meet the severe requirements for efficient and stable monitoring of dynamically changed power systems possessing the steady growth of their size, complexity, and unpredictable behaviors.

As a novel alternative, the latest advanced data driven estimators can assess state evaluation without knowledge of the power network parameters or topology [12-17]. A linearized analysis algorithm was proposed for early event detection using the reduced dimensionality [13]. Lim and DeMarco presented a SVD-based power state evaluation from PMU data, but their methods would be difficult to implement for real time assessment in a large power system due to the high computation burden [14]. Instead of monitoring the raw PMU data, recently, there has been considerable interest in the statistics of PMU measurements. Ghanavati, Hines, and Lakoba sought to identify a statistical state evaluation indicator by calculating the expected variance and autocorrelation of the buses’ voltages and currents [15]. It is noted that the success of these approaches requires an accurate statistical model of measurement noise and load fluctuations. Besides, the constraint that the data dimension should be smaller than the window size also needs to be satisfied in [13,14]. On the other hand, linear eigenvalue statistics (LESs) of the high-dimensional PMU data were utilized for situational awareness or correction analysis.
of the power system in our recent works [12, 15, 17]. Taking advantage of asymptotic properties of high-dimensional random matrix, LES-based methods provided robust power state evaluation using individual window-truncated PMU data. Rather than exploiting individual window-truncated PMU data, this work tries to indicate state evaluation by high-dimensional statistical properties of overall PMU data.

Besides, from the perspective of theoretical research, large deployment of synchronized PMU raises several open issues:

1) How to represent the massive streaming PMU data in the manner of continuous learning of a power system;
2) How to evaluate the real time state evaluation from massive streaming PMU data;
3) Is there any method that can turn the big PMU data into tiny data for the practical use?
4) How to develop a state evaluation estimator without assuming a specific parametric distribution for the data;
5) Does a flexible data driven state evaluation indicator with a wide range of dimensions and sample size exist?

The new metric proposed here is based on multiple high dimensional covariance matrix tests. Tests about high-dimensional covariance matrices have recently increased in popularity. The first attempt on the high-dimensional covariance matrix test presented by Bai and Saranadasa was based on likelihood ratio (LR) test [18]. The LR test works well for normally distributed data on condition that the sample size is larger than the data dimension. Gupta and Xu extended the LR test to non-normal distribution [19] while Bai et al. [20] considered a correction of the CLR test in the case of a wide range of data dimensions. These tests share the basic assumption that the population covariance matrix can be directly substituted by the sample covariance matrix. However, genomic studies showed that such an assumption may not work because these sample covariance matrix based estimators have unnecessary terms which slow down the convergence considerably as the dimension is high [21, 22]. Instead of estimating the population covariance matrix directly, some well-defined distance were proposed to evaluate the difference among populations [23]. Ledoit and Wolf exploited scaled trace-based distance measure between two sub-populations when the data dimension is large compared to the sample size [21]. By exploiting the merits of U-statistics [24], Chen etc. extended the results in [21] to a wide range of data dimension and sample size. However, these works are of high computation burden and focus on the difference of two sub-populations which make them unsuitable for indicating real time state evaluation in massive streaming PMU data.

In this paper, by exploiting the changes in the covariance matrix of different sampling periods of the streaming PMU data, we develop a novel power state evaluation algorithm using the multiple high dimensional covariance matrix tests. The key features of the proposed test statistic are as follows. 1) it can jointly reveal the relative magnitude, duration (or so-called clearing time) and location of a system event; 2) it specifies no parameter distribution of the PMU data, which implies a wide range of the practical applications; 3) it is a real time data driven method without requiring any knowledge of the system model or topology; 4) it is a flexible state evaluation indicator without specifying an explicit relationship between data dimension and sample size; 5) it provides effective computation due to principal component calculation and redundant computation elimination. 6) it implements the asymptotic properties of the high dimensional PMU data to enhance the robustness of the test statistic.

The remainder of this paper is structured as follows. Section I introduces the representation of massive streaming PMU data. Section II presents a power state evaluation approach using multiple high dimension covariance matrix tests. By principal component calculation and redundant computation elimination, an effective calculating method for the proposed test statistic is also developed. In Section IV numerical case studies using synthetic data and real data are provided to evaluate the performance of the proposed state evaluation indicator; The discussion is also included in this section. The conclusion is presented in Section V. For the sake of simplicity, all technical details and some additional case study results are deferred to the Appendices.

II. MASSIVE STREAMING PMU DATA MODELING

It is well accepted that the transient behavior of a large electric power system can be illustrated by a set of differential and algebraic equations (DAEs) as follows [23, 26]:

\[
\dot{x}(t) = f(x(t), u(t), h(t), w) \quad (1)
\]

\[
0 = g(x(t), u(t), h(t), w) \quad (2)
\]

where \(x(t) \in \mathbb{C}^{n \times p}\) are the power state variables, e.g., rotor speeds and the dynamic states of loads, \(u(t)\) represent the system input parameters, \(h(t)\) are the algebraic variables, e.g., bus voltage magnitudes, \(w\) denotes the time-invariant system parameters. \(t \in \mathbb{R}, m, p\) are the sample time, number of system variables, and bus, respectively. The model-based state indicators [2, 11, 15, 25, 27] focus on linearization of nonlinear DAEs in (1) and (2) which gives

\[
\begin{bmatrix}
\Delta \dot{x} \\
\Delta \dot{u}
\end{bmatrix}
= 
\begin{bmatrix}
A & -f_u g_{ux}^{-1} g_{x} \\
0 & -E
\end{bmatrix}
\begin{bmatrix}
\Delta x \\
\Delta u
\end{bmatrix}
+ 
\begin{bmatrix}
0 \\
C
\end{bmatrix}
\xi, \quad (3)
\]

where \(f_x, f_u\) are the Jacobian matrices of \(f\) with respect to \(x, u\) and \(A = f_x - f_u g_{ux}^{-1} g_{x}\). \(E\) is a diagonal matrix whose diagonal entries equal \(t_{cor}\) and \(t_{cor}\) is the correction time of the load fluctuations. \(C\) denotes a diagonal matrix whose diagonal entries are nominal values of the corresponding active \((P)\) or reactive \((Q)\) of loads; \(\xi\) is assumed to be a vector of independent Gaussian random variables.

It is noted that estimating the system state by solving the equation (3) is becoming increasingly more challenging [2, 5] as a consequence of the steady growth of the parameters, say, \(t, p\) and \(m\). Besides, the assumption that \(\xi\) follows Gaussian distribution would restrict the practical application.

As a novel alternative, the lately advanced data driven estimators [12, 17] can assess state evaluation without knowledge of the power network parameters or topology. However, these estimators are based on the analysis of individual window-truncated PMU data. In this work, we seek to provide a method
with the ability of continuous learning of power system from massive streaming PMU data.

We first try to turn the big PMU data (massive high-dimensional PMU data streams) into tiny data (PMU data segments) for practical use. Fig. 1 illustrates the conceptual representation of the structure of the massive streaming PMU data. More specifically, let \( p \) denote the number of the available PMUs across the whole power network, each providing \( m \) measurements. At \( ith \) time sample, a total of \( \kappa = p \times m \) measurements, say \( z_i \), are collected. With respect to each PMU, the \( m \) measurements could contain many categories of variables, such as voltage magnitude, power flow, and frequency, etc. In this work, we develop PMU data analysis assuming each type of measurements is independent. That is, we assert that at each round of analysis, \( \kappa := p \). Given \( q \) time periods of \( T \) seconds with \( K \) Hz sampling frequency in \( ith \) data collection. Let \( n_q = T \times K \) and \( n \) be the window size and sampling number, respectively. A sequence of large random matrix

\[
\left\{ Z_{i1}, Z_{i2}, \ldots, Z_{iq} \right\}_{q \text{ window-truncated data}}, \quad \left\{ Z_{n1}, Z_{n2}, \ldots, Z_{nq} \right\}_{q \text{ window-truncated data}}
\]

(4)

is obtained to represent the collected voltage magnitude data. Note that \( Z_{iq} = \{ z_{i1}, \ldots, z_{inq} \} \), \( i = 1, 2, \ldots, n \).

Fig. 1: Conceptual representation of the structure of the massive streaming PMU data.

In this paper, magnitudes of bus voltages are employed as status data for the following considerations:

1) The voltage magnitudes are one of the most common measurements in power systems and plenty of studies are developed based on them.
2) The voltage magnitude can be collected with no prior knowledge of the topology of power systems. Therefore, we can conduct analysis without knowing the network structures and related parameters.
3) Previous studies \([12, 13, 17]\) have shown that power state indicators based on voltage magnitudes can achieve similar performances compared to other measurements, i.e., currents or power flow.

III. VOLTAGE STABILITY ASSESSMENT

Once we have established the random data flow model for the massive streaming PMU data, the next step is to extract the data analytics. As we all know, power systems are continuously experiencing fluctuations of small magnitudes \([26]\). In the functional setting, it is of interest to test whether or not \( q \) sets of bus voltage curves have similar variations; We assume that the system is initially in a steady-state operating condition for assessing the power state when subjected to a specified disturbance \([26, 28]\). Thus it is interesting to discover the difference of the measurements collected in the normal condition and the abnormal condition by multiple high dimensional covariance matrix tests.

A. Multiple High Dimensional Covariance Matrix Tests

As depicted in the Section \([1]\) a large random matrix flow \( \{ Z_1, Z_2, \ldots, Z_q \} \) is adopted to represent the massive streaming PMU data in one sample period. Instead of analyzing the raw individual window-truncated PMU data \( \{ Z_{yq} \}_{q=1,2,\ldots,q} \) or the statistic of \( Z_y, \) \([12, 15, 17]\), a comprehensive analysis of the statistic of \( \{ Z_1, Z_2, \ldots, Z_q \} \) is conducted in the following.

More specially, with \( \Sigma_i \) as the covariance matrix of \( ith \) collected PMU measurements, we test the hypothesis:

\[
H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_q \quad H_1 : \exists \ j, k \text{ s.t. } \Sigma_j \neq \Sigma_k.
\]

(5)

It is worthy noting that the hypothesis \( 5 \) is a famous testing hypothesis in multivariate statistical analysis \([29, \text{Chapter 10}]\) which aims to study samples that share or approximately share some of the same distributions and consider using a set of samples (data streams denote in equation (4) in this paper), one from each population, to test the hypothesis that the covariance matrices of these populations are equal.

B. Proposed test statistic

The LR test \([18]\) and CLR test \([20]\) as introduced in the Section \([1]\) are the most commonly test statistics for the hypothesis in \( 5 \). For the readers’ convenience, we briefly explain the technical details in the Appendix A. These tests can be understood by replacing the population covariance matrix \( \Sigma_q \) by its sample covariance matrix \( Y_q \). Direct substitution of \( \Sigma_q \) by \( Y_q \) brings invariance and good testing properties as shown in \([18]\) for normally distributed data. The test statistic \( V_2 \) may not work for high-dimensional data as demonstrated in \([21, 22]\). Besides, the estimator \( V_3 \) has unnecessary terms, which slows down the convergence considerably when dimension of PMU data is high \([22, 23]\). In such situations, to reduce the drawbacks, trace criterion \([22]\) is more suitable to the testing problem. Instead of estimating the population covariance matrix directly, a well defined distance measure exploiting the difference among data flow \( \{ Z_1, Z_2, \ldots, Z_q \} \) is conducted, that is, the trace-based distance measure between \( \Sigma_s \) and \( \Sigma_t \) is

\[
\text{tr} \left\{ (\Sigma_s - \Sigma_t)^T \right\} = \text{tr} (\Sigma_s^2) + \text{tr} (\Sigma_t^2) - 2\text{tr} (\Sigma_s \Sigma_t),
\]

(6)

where \( \text{tr} (\cdot) \) is the trace operator. Instead of estimating \( \text{tr} (\Sigma_s^2), \) \( \text{tr} (\Sigma_t^2) \), and \( \text{tr} (\Sigma_s \Sigma_t) \) by sample covariance matrix based estimators, we develop an unbiased estimator according to the U-statistics \([24]\) which allow a minimum-variance unbiased
estimator to be derived from each unbiased estimator of an estimable parameter for large classes of probability distributions. Specially, for \( l = \{s, t\} \in \Omega = \{1 \leq s, t \leq q, s \neq t\}, \)

\[
A_t = \frac{1}{n_g (n_g - 1)} \sum_{i,j} (Z_i'Z_j)^2 - \frac{2}{n_g (n_g - 1) (n_g - 2)} \sum_{i,j,k} Z_i'Z_jZ_k \sum_{i,j,k} Z_i'Z_jZ_k - \frac{1}{n_g (n_g - 1) (n_g - 2) (n_g - 3)} \sum_{i,j,k,h} Z_i'Z_jZ_k Z_h.
\]

(7)

is proposed to estimate \( \text{tr} \left( \Sigma^2 \right) \). It is noted that \( \sum^* \) represents summation over mutually distinct indices. For example, \( \sum^*_{i,j,k} \) says summation over the set \( \{(i,j,k): i \neq j, j \neq k, k \neq i\} \). Similarly, the estimator for \( \text{tr} \left( \Sigma_s \Sigma_t \right) \) can be expressed as

\[
C_{st} = \frac{1}{n_g^2} \sum_{i,j} (Z_i'Z_j)^2 - \frac{1}{n_g^2} \sum_{i,j} (Z_i'Z_j) \sum_{i,j} (Z_i'Z_j) - \frac{1}{n_g^2} \sum_{i,j,k,h} (Z_i'Z_j) (Z_kZ_h).
\]

(8)

The test statistic that measures the distance between \( \Sigma_s \) and \( \Sigma_t \) is

\[
V_{st} = A_s + A_t - C_{st}.
\]

Then the proposed test statistic can be expressed as:

\[
V_1 = \frac{1}{q(q-1)} \sum_{\{s,t\} \in \Omega} V_{st}.
\]

(10)

As \( p, n_g \to \infty \), the asymptotic normality of the test statistic \( V_1 \) is presented in the following:

**Theorem III.1.** Let \( \sigma^2_{st} = \frac{1}{n_g} (A_s + A_t) \). Assuming the following conditions:

1. For any \( k \) and \( l \in \{s, t\}, \text{tr} \left( \Sigma_k \Sigma_i \right) \to \infty \) and \( \text{tr} \left( \left\{ \Sigma_i, \Sigma_j \right\} \right) = O \left( \text{tr} \left( \Sigma_k \Sigma_i \right) \right) \).

2. For \( i = 1, 2, \ldots, n_q \), \( Z^{(i)} \) are independent and identically distributed \( p \)-dimensional vectors with finite 8th moment.

Under above conditions,

\[
L = \frac{V_{st}}{\sigma_{st}} \overset{d}{\to} N(0, 1).
\]

**Proposition III.2.** For any \( q \geq 2 \), as \( p, n_g \to \infty \), the proposed test statistic \( V_1 \) satisfies

\[
V_1 \overset{d}{\to} N(\mu, \sigma^2),
\]

where \( \mu \approx 0, \sigma^2 = \sum^* \sigma^2_{st} \).

Let \( R = \frac{V_1}{\sigma_{V_1}} \), the false alarm rate (FAR) for the proposed test statistic can be represented as

\[
P_{FAR} = P(R > \alpha|H_0) = \int_R^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left(- \frac{t^2}{2} \right) dt = Q(R),
\]

where \( Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left(- \frac{t^2}{2} \right) dt \). For a desired FAR \( \tau \), the associated threshold should be chosen such that

\[
\alpha = Q^{-1}(\tau).
\]

Otherwise, the detection rate (DR) can be denoted as

\[
P_{DR} = P(R \geq Q(\alpha)|H_1).
\]

It is noted that the computation complexity of proposed test statistic in (11) is \( O(n_g^4) \), which limits its practical application. Here, we propose an effective approach to reducing complexity of the proposed test statistic from \( O(n_g^4) \) to \( O(n_g^2) \) by principal component calculation and redundant computation elimination. For simplicity, the technical details are deferred to the Appendix B.

**C. Continuous Learning of the Power System**

Based on the proposed multiple high-dimensional test in Section III-B, we propose a method in the continuous manner to indicate the state evaluation. Details are shown in the following:

Let

\[
T_{trn} = \left[ T^{1\text{trn}}_{\text{trn}}, \ldots, T^{q\text{trn}}_{\text{trn}} \right]
\]

be the total training period. It is presumed that the power system is under normal operation during time period \( T_{trn} \).

For \( i = 1, 2, \ldots, n \), the collected PMU data flow \( \{Z_{1i}, Z_{2i}, \ldots, Z_{qi}\} \) is employed for continuous learning of the power system parameters, namely, mean and variance of the proposed test statistic, detection threshold in (12), and then power system state. Specifically,

1. **Estimating the relative magnitude and duration of the system event:** Using the proposed test statistic in (10), a system event can be identified with several samples of PMU data when the system event indicator satisfies

\[
|V_1 - \mu| \geq \gamma,
\]

(14)

where \( \mu = 3\sigma \) are the system-dependent parameters which can be learned from explanatory historical PMU data in the training procedure. The relative magnitude of a system event equals the test statistic \( V_1 \). Given that a system event occurs in sample period \( T_{\text{test}} \), for \( j = 1, 2, \ldots, c \), denotes the test data flow as \( \{Z_{11}, Z_{22}, \ldots, Z_{cq}\} \), the duration of the event can be roughly estimated by

\[
T_{\text{dur}} = \sum_{j=1}^{c} q * T * \omega_j,
\]

(15)

where

\[
\omega_j = \begin{cases} 
1, & |V_1 - \mu| \geq \gamma \\
0, & |V_1 - \mu| < \gamma 
\end{cases}
\]
2) Determination of the most sensitive PMU: According to the data analysis in [III-C1] the voltage event addressed on a power system can be identified. Then, the determination of the most sensitive PMU is another important part to be investigated with respect to a system event.

The fact that every fault has its own effect on a power system [26] stimulates us to find the location of most sensitive PMU. According to the data analysis in Section [III-C1] we are able to determine the time when a system event occurs, say, $T_1$. Assume that the power system operates under normal condition during the time period of $T_1-1$ and there are $p$ types of influential factors during a sampling time $T_1$. Denote $Z^{(i)} = \{Z_{i1}, Z_{i2}, \ldots, Z_{iq}\}$, $Z^{(j)} = \{Z_{j1}, Z_{j2}, \ldots, Z_{jq}\}$ and $Z^{(k)} = \{Z_{k1}, Z_{k2}, \ldots, Z_{kp}\}$ as the PMU data flow collected during sample time $T_1-2$, $T_1-1$ and $T_1$. For $l = 1, 2, \ldots, p$, the measured data of each factor are formed as a row vector $c_l^{(T_1)}$. In order to reveal the most sensitive PMU, we form a factor matrix by duplicating $\kappa$ times for each factor $c_l^{(T_1)}$, say,

$$C^{(T_1)} = \begin{bmatrix} c_1^{(T_1)} \\ \vdots \\ c_p^{(T_1)} \end{bmatrix}_{\kappa \times N} ,$$

(16)

where the parameter $N = q \cdot n_g$, $\kappa = r \cdot \log p$ and $r$ is the rank of $Z^{(i)}$. For $l = 1, 2, \ldots, p$, we can construct two expansion matrices for parallel data analysis, formulated by

$$A_1^{(l)} = C^{(T_1)} \cdot Z^{(i)} , \quad A_2^{(l)} = C^{(T_1)} \cdot Z^{(j)} .$$

(17)

Substitute data flows $A_{1l}$ and $A_{2l}$ into the test statistic in [10], the location of most sensitive PMU data (denoted as $loc$) during the sample time $T_1$ can be expressed as

$$loc = \max \left( \frac{1}{l=1,2,\ldots,p} \left( V_1^{(l)} \right) \right) ,$$

(18)

where $\max (x_j) = j$.

For the readers’ convenience, the technological process of the proposed test statistic for power state evaluation are summarized in the following:

| Implementation of the proposed state evaluation indicator. |
|-------------------------------------------------------------|
| 1a): Off-line training period (System-dependent parameters learning): |
| 1b): collect the PMU data and represent them using [10]; |
| 1c): calculate mean and variance of the proposed test statistic; |
| 1d): determine the event indicator threshold $\gamma$ using [14]; |
| 2a): Online power state indicating: |
| 2b): acquire the test data flow: $Z_{j1}, \ldots, Z_{jq}$, $j = 1, 2, \ldots, c$; |
| 2c): determine whether there is an event using [14]: |
| if no event detected: |
| \quad add the test data flow into history data; |
| \quad go back to the step 1a); |
| else: |
| \quad go to step 2d); |
| 2d): Determine the relative magnitude, duration and location of the system event using [10], [15] and [18], respectively; |
| 3a): FAR [12] and DR [13] analysis; |
| 3b): the effect of measurement noise analysis; |
| 3c): the effect of parameter $q$ analysis. |

So far, the power state evaluation by the proposed test statistic is established. Case studies to evaluate the practical performance of the proposed test statistic will be depicted in detail in the following section.

IV. CASE STUDIES AND DISCUSSIONS

The proposed test statistic for power state evaluation are numerically evaluated by the power network benchmarks, namely the IEEE 30-, 118-bus system, a Polish 2383-bus system [31], and a real 34-PMU system. For the synthetic data, the admittance matrices and the underlying power system states are generated by MATPOWER package [32]. It is noted that the measurement noise is simulated as uncorrelated Gaussian or Gama distribution with a standard deviation per component of 0.05 for voltages [26, 32]. We report results from case studies which are designed to evaluate the performance of the proposed test for the power state evaluation in the following.

A. Effect of measurement noise on the power state evaluation

As discussed in Section [I] PMUs offer highly accurate measurements (i.e., voltage, current phasors and frequency) when operating under steady-state conditions. Significant error creeps into the measurements, however, while operating under transient conditions. The amount of measurement error varies from one manufacturer to another due to the difference in the method used to calculate the output quantities. This poses a serious question to the power state indicator’s ability to monitor dynamics of a power system at the time of disturbances when we have no prior knowledge about measurement noise. So the effect of measurement noise on the power state evaluation is firstly studied as follows.

Assuming that the power system operates under normal state, we first investigate the effect of measurement noise and window size on the state evaluation using synthetic data. With respect to the proposed test in [10], we generate $p$-dimensional data independent multivariate data models using the linearized measurement model in [3]. Let $z_0$ be the initial state of the power system. The nominal significance level [20] of the data and parameter $q$ are set to 5% and 10, respectively. For $i = 1, 2, \ldots, n_g$, we consider two scenarios regarding the innovation random vector $z_i^{(i)}$:

1) $z_i^{(i)}$ are $p$-dimensional normal random vector with mean $z_0$ and variance diag $\{0.05z_0\}$.

2) $z_i^{(i)} = \begin{bmatrix} z_1^{(i)} \\ \vdots \\ z_p^{(i)} \end{bmatrix}$ consist of independent random variables $z_j^{(i)}$ which are standard Gamma($z_0, 0.2236$) + 0.7764$z_0$ random variables.

It is noted that the proposed test statistic imposes no restriction on the relationship between the data dimension and sample size. To mimic the buses deployed in the power system, we have $p \in \{30, 118, 2383\}$. A wide range of sample window sizes are denoted as $n_s \in \{30, 100, 300, 1000, 2500\}$. The simulation results (Tab. [I] and Tab. [II] reported in this section are based on 1000 independent Monte Carlo simulations.

The simulation results in Tab. [I] and Tab. [II] show that DR of the covered test statistics increase as the dimension and sample sizes become larger. Many entries of the DR of the tests
TABLE I: DR and FAR of the test statistics with GSN.

| $(p, n_g, q)$ | LR test | CLR test | Proposed test |
|--------------|---------|----------|---------------|
|              | DR      | FAR      | DR            | FAR     | DR      | FAR     |
| (30,30,10)   | 0.595   | 0.059    | 0.651         | 0.067   | 0.694   | 0.061   |
| (30,100,10)  | 0.742   | 0.064    | 0.899         | 0.061   | 0.912   | 0.058   |
| (30,300,10)  | 0.901   | 0.089    | 0.955         | 0.057   | 0.979   | 0.047   |
| (30,1000,10) | 0.958   | 0.134    | 0.997         | 0.054   | 0.999   | 0.039   |
| (2383,2500,10) | 1      | 0.296    | 1             | 0.049   | 1       | 0.049   |
| (118,30,10)  | -       | -        | 0.924         | 0.047   | 0.985   | 0.059   |
| (118,100,10) | -       | -        | 0.957         | 0.051   | 0.993   | 0.055   |
| (118,300,10) | 0.995   | 0.149    | 0.993         | 0.053   | 1       | 0.049   |
| (118,1000,10)| 1       | 0.390    | 1             | 0.048   | 1       | 0.045   |
| (118,2500,10)| 1       | 0.483    | 1             | 0.045   | 1       | 0.043   |
| (2383,30,10) | -       | -        | 0.991         | 0.063   | 0.995   | 0.058   |
| (2383,100,10)| -       | -        | 1             | 0.055   | 1       | 0.053   |
| (2383,300,10)| -       | -        | 1             | 0.051   | 1       | 0.050   |
| (2383,1000,10)| -   | -       | 1             | 0.046   | 1       | 0.048   |
| (2383,2500,10)| 1 | 0.891   | 1             | 0.047   | 1       | 0.049   |

TABLE II: DR and FAR of the test statistics with GMN.

| $(p, n_g, q)$ | LR test | CLR test | Proposed test |
|--------------|---------|----------|---------------|
|              | DR      | FAR      | DR            | FAR     | DR      | FAR     |
| (30,30,10)   | 0.471   | 0.067    | 0.553         | 0.073   | 0.476   | 0.069   |
| (30,100,10)  | 0.660   | 0.163    | 0.643         | 0.075   | 0.775   | 0.067   |
| (30,300,10)  | 0.791   | 0.289    | 0.816         | 0.067   | 0.891   | 0.067   |
| (30,1000,10) | 0.958   | 0.334    | 0.894         | 0.060   | 0.953   | 0.063   |
| (30,2500,10) | 0.996   | 0.596    | 0.934         | 0.057   | 0.989   | 0.055   |
| (118,30,10)  | -       | -        | 0.801         | 0.066   | 0.885   | 0.059   |
| (118,100,10) | -       | -        | 0.879         | 0.059   | 0.967   | 0.063   |
| (118,300,10) | 0.932   | 0.349    | 0.942         | 0.063   | 0.995   | 0.056   |
| (118,1000,10)| 0.999   | 0.575    | 0.970         | 0.056   | 1       | 0.052   |
| (118,2500,10)| 1       | 0.977    | 0.998         | 0.051   | 1       | 0.055   |
| (2383,30,10) | -       | -        | 0.947         | 0.062   | 0.984   | 0.061   |
| (2383,100,10)| -       | -        | 0.983         | 0.058   | 0.999   | 0.060   |
| (2383,300,10)| -       | -        | 1             | 0.059   | 1       | 0.054   |
| (2383,1000,10)| -   | -       | 1             | 0.049   | 1       | 0.052   |
| (2383,2500,10)| 1 | 1       | 1             | 0.046   | 1       | 0.048   |

approach 1 in both of the scenarios of Gaussian distributed noise (GSN) and Gamma distributed noise (GMN). Besides, as shown in the Table I and Table II, FAR of the proposed test converge to the nominal 5% quite rapidly with $p$ and $n_g$ increase for both GSN and GMN; Meanwhile, the convergence of the FAR to the nominal level for GSN is slower than GMN. On the other hand, the LR test is not applicable for $p \geq n_g$ and CLR test shows slower convergence than the proposed test. In other words, the proposed test statistic has more accurate DR and robust FAR in a quite wide range of dimensionality and distributions while the LR test and the CLR test are vulnerable to variation in the data dimension and noise distribution. This could be understood as the proposed test is both asymptotic and nonparametric.

B. Effect of the parameter $q$ on the power state evaluation

As depicted in the Section II, the parameter $q$ is an important factor for state evaluation. More details illustrated by experimental data are shown in the following. We fix the total data size as 600, that is $q \times n_g = 600$ in first experiment while setting the window size $n_g$ as 100 in the second one. Two kinds of measurement noise shown in the Section IV-A are considered. It is noted that the notations $gauss-30$ and $gama-30$ mean that the measurement noise adopted are GSN and GMN with the number of PMUs at $p = 30$, respectively. Similar definitions also work for other notations, i.e., $gauss-118$, $gama-118$, $gauss-2383$, and $gama-2383$.

![Fig. 2a](image1.png)  ![Fig. 2b](image2.png)

Fig. 2: Effect of the parameter $q$ on the power state evaluation.

Fig. 2a shows that the DR decreases as $q$ increases for the first experiment while Fig. 2b illustrates that the DR shows a positive response to the increase of $q$ in the second one. The selection of a medium size $q$ is the trade-off between the DR and the realtime performance. In the rest of the experiments, we set the parameter $q$ as $q = 5$.

C. Online Power State Evaluation Using the Synthetic Data

The performance of the power state evaluation using the proposed test statistic is evaluated by the simulated data generated from IEEE 30-bus, IEEE 118-bus, and a polish 2383-bus system, respectively. The specific details of the systems are referred to the case30.m, case118.m and case2383.m in the Matpower package and Matpower 5.1-User’s Manual [33]. In the simulations, changes on the active load of each bus are considered as potential factors. Besides, each change of a factor is described as a signal. Three kinds of signals that affect the operating state of the test system are considered. For simplicity, the signals for each factor are shown in Tab. III IV and V. $\rho$ denotes the number of P-V nodes in the test systems and is chosen on a random basis. For the sake of simplicity, the case studies based on the IEEE 118-bus system are presented below. The results generated from IEEE 30-bus system and the Polish 2383-bus system are deferred to the Appendix C.

Table: Signal Type I: Voltage Dip

| Bus | Duration | Active Load (MW) |
|-----|----------|-----------------|
|     | $t = 1 \sim 300$ | 40.0         |
| $\rho$ | $t = 301 \sim 600$ | 80.0         |
| $t = 601 \sim 1000$ | 120.0         |
| Others | $t = 1 \sim 1000$ | Unchanged |

The signals are generated in load of P-V node $\rho = 63$ for the case of IEEE 118-bus system. During the training period, 5 minutes of data are collected when the system is under normal condition. Let $p = 118$, $n_g = 100$, $q = 5$. Two kinds of measurement noise shown in the Section IV-A are considered. As shown in the Section III-B, the proposed test statistic satisfies $A \overset{\text{d}}{\rightarrow} \mathcal{N}(0, 1)$. The theoretical bound in Fig.
The probability density function (PDF) of $\lambda$. Fig. 3 shows that the mean and variance of $\lambda$ fit fabulously with theoretical ones.

![Parameter learning with GSN.](a) ![Parameter learning with GMN.](b) (a) Parameter learning with GSN. (b) Parameter learning with GMN. Fig. 3: Parameter learning of IEEE 118-bus system

The power state evaluation begins at 301th s, 60 seconds of data are collected. Three kinds of system events are generated in load of 63th bus from 320s to 340s, respectively. According to the results in Fig. 3 and event indicators (14) and (15), we can know that the event occurs at 301s and the actual duration of the signals can be calculated as $t_{dur} = 1000/(q \times n_g) \times 10 = 20s$. Based on the above analysis, we can then determine the location of the most sensitive bus using (18). The results in Fig. 4 demonstrate that the 63th bus is the most sensitive bus in presence of all three kinds system events when the measurement noise is set as GSN or GMN.

![Parameter learning with GSN.](a) ![Parameter learning with GMN.](b) (a) Parameter learning with GSN. (b) Parameter learning with GMN. Fig. 3: Parameter learning of IEEE 118-bus system

D. A Real Data Analysis

In this section, we evaluate the efficacy of the proposed test statistic for the power system state. For the experiments shown in the following, the real power flow is of a chain-reaction fault that happened in the China power grids in 2013. The PMU number, the sample rate, and the total sample time are $p = 34$, $K = 50Hz$ and 284s, respectively. The chain-reaction fault happened from $t = 65.4s$ to $t = 73.3s$.

Let $q = 5$, $n_g = 50$. Fig. 4 shows that the mean and variance of $\lambda$ agree well with theoretical ones. Based on the results in Fig. 4 and event indicators (14) and (15), the occurrence time

and the actual duration of the event can be identified as $t_0 = 65s$ and $t_{dur} \approx 8s$, respectively. Similar to the data analysis above, we can then determine the location of the most sensitive bus using (18). The results illustrate that 17th and 18th PMU are the most sensitive PMUs, which are in accordance with the actual accident situation.

E. More Discussions in Details

The general objective of the wide deployment of the PMU installation equipments is to eventually make a transition from the conventional supervisory control and data acquisition based measurement system to a more superior measurement system that will use synchronized measurements collected from geographically distant locations and increase the power state evaluation by monitoring a wide area of the power system in real time. The main problem raised from large power systems is how to tackle the tough situation when massive large dimensional measurements get together. This paper enhances the connection between big data analysis and large scale power system, by providing a comprehensive analysis for massive streaming PMU data. In the case studies

| Bus | Duration | Active Load (MW) |
|-----|----------|----------------|
| $t = 1 \sim 300$ | -10.0 |
| $t = 301 \sim 540$ | -25.1 |
| $\rho$ | $t = 541 \sim 780$ | -39.3 |
| $\rho$ | $t = 781 \sim 900$ | -62.7 |
| $t = 901 \sim 1000$ | -75.3 |
| Others | $t = 1 \sim 1000$ | Unchanged |

| Bus | Duration | Active Load (MW) |
|-----|----------|----------------|
| $t = 1 \sim 300$ | 10.0 |
| $t = 301 \sim 600$ | 60.0 |
| $\rho$ | $t = 601 \sim 900$ | 120.0 |
| $\rho$ | $t = 901 \sim 1000$ | 35.0 |
| Others | $t = 1 \sim 1000$ | Unchanged |

TABLE IV: Signal Type II: Voltage Swell

| Bus | Duration | Active Load (MW) |
|-----|----------|----------------|
| $t = 1 \sim 300$ | 10.0 |
| $t = 301 \sim 600$ | 60.0 |
| $\rho$ | $t = 601 \sim 900$ | 120.0 |
| $\rho$ | $t = 901 \sim 1000$ | 35.0 |
| Others | $t = 1 \sim 1000$ | Unchanged |

TABLE V: Signal Type III: Voltage Dip and Swell

| Bus | Duration | Active Load (MW) |
|-----|----------|----------------|
| $t = 1 \sim 300$ | 10.0 |
| $t = 301 \sim 600$ | 60.0 |
| $\rho$ | $t = 601 \sim 900$ | 120.0 |
| $\rho$ | $t = 901 \sim 1000$ | 35.0 |
| Others | $t = 1 \sim 1000$ | Unchanged |

| Bus | Duration | Active Load (MW) |
|-----|----------|----------------|
| $t = 1 \sim 300$ | 10.0 |
| $t = 301 \sim 600$ | 60.0 |
| $\rho$ | $t = 601 \sim 900$ | 120.0 |
| $\rho$ | $t = 901 \sim 1000$ | 35.0 |
| Others | $t = 1 \sim 1000$ | Unchanged |
above, we demonstrate the efficacy of the proposed power state indicator, including the parameter selection and power state indication, under different scenarios. However, there are still some interesting details left.

Results in Section IV-A show that our proposed non-parameter power state indicator work well in two types of measurement noise (GSN and GMN); More types of noise are also suggested and studied. Besides, as shown in Section II and Section IV-B, the parameter \( q \) acts as an important role in power state evaluation. There should be some connections between phenomena in power systems and \( q \) in mathematics. Due to the limit space, we leave that for our future work. Moreover, three kinds of signals that affect the operating state of the test system are considered in Section IV-C. In a real power system, those signals are mainly caused by the three types of system events: control input changes, initial condition changes or system topology changes \([20]\). Those signals would occur irregularly in the modern power system where enormous new energy resources and renewable energy sources exist. The results in Section IV-C demonstrate that our advanced power state indicator could reveal the relative magnitude, duration and location of any system event. More detailed case studies need to be developed to harvest the maximum benefit of the proposed power state indicator for monitoring, protection, and control of a power system under all operating conditions (i.e., the scenario where there exist more types of signals or more complicated signals).

As mentioned in Section II, the new metric proposed in this paper is based on multiple high dimensional covariance matrix tests. The advantage of the asymptotic property of the proposed test statistic is studied and employed. The results in Fig. 3 and Fig. 5 show that the p.d.f of the proposed test statistic fits fabulously within the theoretical bound for synthetic data while the performance is a little bit worse for real data. Therefore, deducing a more precise bound for the proposed test statistic is necessary in future work.

This work develops an essential mathematical framework with a sound theoretical guarantee. Besides, the favorable properties (i.e., parameter-independence, flexibility and low computation complexity) of the proposed method would benefit not only readers with electric backgrounds but a large number of interested readers (i.e., from computer science, bio-engineering, or applied mathematics) who encounter similar problems.

V. Conclusion

Motivated by the immediate demands of big data analysis for large scale smart grid, this paper proposes a real time data-driven method to indicate the state evaluation from massive streaming PMU data. First, we represent the PMU data as a sequence of large random matrices. This is a crucial part for the power state evaluation as it turns the big PMU data into tiny data for practical use. Rather than employing the raw PMU data or window-truncated data, a comprehensive analysis of PMU data flow, namely, the multiple high-dimensional test, is then proposed to indicate the power system state. The case studies based on synthetic data and real data show that the proposed test statistic can accurately reveal the relative magnitude, duration, and location of a system event in polynomial time.

The current work provides a fundamental exploration of data analysis for massive streaming PMU data. Much more attention is to be paid along this research direction, such as classification of power events from massive streaming PMU data. It is also noted that this work is a data-driven method which is a new substitute for power system state estimation. The combination of power system scenario analysis and the data driven methods is encouraged to be further investigated for a better understanding of the power system state.

VI. Acknowledgement

The authors would like to thank the anonymous reviewers for a critical reading of the manuscript, which led to a significant improvement of this paper.
APPENDIX A
THE TRADITIONAL HIGH-DIMENSIONAL COVARIANCE MATRIX TESTS

For the readers’ convenience, we give a brief description of the likelihood ratio (LR) test statistic and correction of the likelihood ratio (CLR) test statistic in the following.

Let \( \sum_{g=1}^{q} n_g = n \) be the total sample size, \( \bar{z}_g = \sum_{k=1}^{n_g} z_{gk} \),

\[
Y_g = \frac{1}{n_g - 1} \sum_{k=1}^{n_g} (z_{gk} - \bar{z}_g)(z_{gk} - \bar{z}_g)',
\]

\[
Y = \sum_{g=1}^{q} Y_g.
\]

The LR test [18] for testing hypothesis equation (5) is

\[
V_2 = \frac{\prod_{g=1}^{q} |Y_g|^{\frac{N_g}{2}}}{|Y|^{\frac{1}{2}M}},
\]

where

\[
N_g = n_g - 1, \quad M = N_1 + N_2 + \cdots + N_q = n - q.
\]

It is noted that the calculation of the numerator and denominator of \( V_2 \) will lead to overflow as \( n_g \) becomes large. To overcome the overflow difficulty, a CLR test [20] for testing the equality of more population covariance matrices is shown as follows. Let

\[
V_{2h} = \frac{\prod_{g=1}^{q} |Y_g|^{\frac{N_g}{2}}}{|Y|^{\frac{1}{2}M}},
\]

where \( h = 2, 3, \cdots, q \). Then \( V_2 = \prod_{h=2}^{q} V_{2h} \).

The CLR test statistic is

\[
V_3 = \sum_{h=2}^{q} \frac{2}{N_1 + N_2 + \cdots + N_{h-1}} \log T_{1h} - p f (y_{1h}, y_{2h}),
\]

where

\[
y_{1h} = \frac{p}{N_1 + N_2 + \cdots + N_{h-1}}, \quad y_{2h} = \frac{p}{N_h},
\]

and

\[
f (y_1, y_2) = \frac{y_1 + y_2 - y_1 y_2 \log \left( \frac{y_1 + y_2}{y_1 y_2} \right)}{y_1 y_2} + \frac{y_1 (1 - y_2) \log (1 - y_2) + y_2 (1 - y_1) \log (1 - y_1)}{y_1 y_2 (1 - y_1)}
- \frac{y_1}{y_1 + y_2} \log \frac{y_1}{y_1 + y_2} - \frac{y_2}{y_1 + y_2} \log \frac{y_2}{y_1 + y_2}.
\]

APPENDIX B
COMPUTATION ASPECT OF THE PROPOSED TEST STATISTIC

From Equation (7)-(9), we know that the computational complexity of calculating the test statistics \( A_s, A_t \) and \( C_{st} \) are \( O(\varepsilon_1 n_g^2) \), \( O(\varepsilon_2 n_g^2) \) and \( O(\varepsilon_3 n_g^2) \), respectively. With the increasing scale of PMU deployment and the increasing complexity of issues addressed by it, which is a newly raised challenge for the power state evaluation and quality control of a power system. Here, we propose a lower complexity method to calculate \( A_s, A_t, \) and \( C_{st} \) by redundant computation elimination and principal component calculation. Technical details are elaborated in the following.

A. Redundant Computation Elimination

We first consider eliminating the index-wise redundant computation during calculating the term \( A_{l_{i\in\{s,t\}}} \). Let

\[
A_{l_1} = \frac{1}{n_g (n_g - 1)} \sum_{i \neq j} (z'_{l_1} z_{l_{ij}})^2,
\]

\[
A_{l_2} = \frac{2}{n_g (n_g - 1) (n_g - 2)} \sum_{i,j,k} z'_{l_1} z'_{l_{ij}} z_{l_{ijk}},
\]

and

\[
A_{l_3} = \frac{1}{n_g (n_g - 1) (n_g - 2) (n_g - 3)} \sum_{i,j,k,h} z'_{l_1} z'_{l_{ij}} z_{l_{ijk}} z_{l_{l_{ijh}}},
\]

It is easy to find that indices \( i, j, k, l \) in \( A_{l_1}, A_{l_2} \) and \( A_{l_3} \) are invariant with respect to the swapping places. Let

\[
\Omega_1 = \{ \{i, j\} : 1 \leq i, j \leq n_g, i \neq j \},
\]

\[
\Omega_2 = \{ \{i, j, k\} : 1 \leq i, j, k \leq n_g, i \neq j \neq k \},
\]

\[
\Omega_3 = \{ \{i, j, k, h\} : 1 \leq i, j, k, h \leq n_g, i \neq j \neq k \neq h \}.
\]

Specially, we are to determine unrepeated sets of the indices from \( \Omega_1, \Omega_2 \) and \( \Omega_3 \) when calculating \( A_{l_1}, A_{l_2} \) and \( A_{l_3} \). Following the permutations and combinations principle in [34], the unrepeated ensembles can be expressed as

\[
\hat{\Omega}_1 = \{ \{i, j\} : 2 \leq i \leq n_g, 1 \leq j \leq i - 1 \},
\]

\[
\hat{\Omega}_2 = \{ \{i, j, k\} : 3 \leq i \leq n_g, 2 \leq j \leq i - 1, 1 \leq k \leq j - 1 \},
\]

\[
\hat{\Omega}_3 = \{ \{i, j, k, h\} : |\{4 \leq i \leq n_g, 3 \leq j \leq i - 1\} \cup \{2 \leq k \leq j - 1, 1 \leq h \leq k - 1\} | \}.
\]

Let \( Q_{n_g}^r = n_g!/(n_g - r)! \). Then, \( A_{l_1}, A_{l_2} \) and \( A_{l_3} \) can be expressed by

\[
A_{l_1} = \frac{2}{Q_{n_g}^r} \sum_{\{i, j\} \in \hat{\Omega}_1} (z'_{l_1} z_{l_{ij}})^2,
\]

\[
A_{l_2} = \frac{6}{Q_{n_g}^r} \sum_{\{i, j, k\} \in \hat{\Omega}_2} z'_{l_1} z_{l_{ij}} z_{l_{ijk}},
\]

and

\[
A_{l_3} = \frac{24}{Q_{n_g}^r} \sum_{\{i, j, k, h\} \in \hat{\Omega}_3} z'_{l_1} z_{l_{ij}} z_{l_{ijk}} z_{l_{l_{ijh}}}.
\]

As a result, the computational complexity of calculating \( A_{l_1}, A_{l_2}, \) and \( A_{l_3} \) is reduced by a factor of \( 1/2, 1/6, \) and \( 1/24 \) compared with direct manipulation, respectively.

Besides, we notice that manipulation of \( A_{l_1}, A_{l_2}, \) and \( A_{l_3} \) is completed in sequence and this manipulation is inefficient because of the repeated vector multiplication operations. For instance, vector multiplication \( z'_{l_1} z_{l_{ij}} \) is repeated many times when calculating \( A_{l_1}, A_{l_2} \) and \( A_{l_3} \). This kind of repeated calculation can be avoided by the following steps.
Let $Z^i_j$ be voltage-relevant matrix whose elements are

$$Z^i_j = z^i_j, \{i, j\} \in \Omega_1.$$  

Then $A_{11}$, $A_{12}$ and $A_{13}$ can be equivalently denoted as

$$A_{11} = \frac{2}{Q_n^2} \sum_{\{i, j\} \in \Omega_1} (z^i_j)^2,$$

$$A_{12} = \frac{6}{Q_n^4} \sum_{\{i, j, k\} \in \Omega_2} Z^i_j Z^j_k,$$

and

$$A_{13} = \frac{24}{Q_n^4} \sum_{\{i, j, k, h\} \in \Omega_3} Z^i_j Z^j_k.$$  

The aforementioned equivalent expression means that we can compute $z^i_j, z^i_j$ only once during the progress in calculating $A_{11}$, $A_{12}$ and $A_{13}$. Thus the computing time is reduced to $1/n^2_4$ of the conventional calculation of $A_{12}$ and $A_{13}$.

Similarly, the computation burden of calculating $C_{s,t}$ can be alleviated by repeating the steps above. Here we only provide the result. $C_{s,t}$ can be equivalently denoted as

$$C_{s,t} = \frac{2}{Q_n^2} \sum_{\{i, j\} \in \Omega_1} (Y_{ij})^2 - \frac{12}{Q_n^2} \sum_{\{i, j, k\} \in \Omega_2} (Y_{ij} Y_{jk} + Y_{ij} Y_{jk}) + \frac{24}{Q_n^4} \sum_{\{i, j, k, h\} \in \Omega_3} Y_{ij} Y_{kh},$$

where $Y_{ij} = z^i_j z^i_j, \{i, j\} \in \Omega_1$.

### B. Principal Component Calculation

Let $A = B + C$, where $A, B, C$ are positive random variables. Let $n_g$ be a large positive number, say, 100. If the condition that $C/B < 1/n_g$ is satisfied, then $B$ is called the principal component of $A$. Then we introduce the principal component calculation.

It can be noted that the magnitude of voltage measurements are positive, that is,

$$Z^i_j > 0, \{i, j\} \in \{\Omega_1 \cup \Omega_2 \cup \Omega_3\},$$

then

$$\sum_{\{i, j\} \in \Omega_1} (Z^i_j)^2 > \sum_{\{i, j, k\} \in \Omega_2} Z^i_j Z^j_k > \sum_{\{i, j, k, h\} \in \Omega_3} Z^i_j Z^j_k.$$  

For $n_g \gg 1$, divide $A_{12}$ and $A_{13}$ by $A_{11}$, respectively, we can get

$$A_{12} = \frac{6}{Q_n^2} \sum_{\{i, j, k\} \in \Omega_2} (Z^i_j)^2 < \frac{3}{n_g-2} < 1,$$

and

$$A_{13} = \frac{24}{Q_n^4} \sum_{\{i, j, k, h\} \in \Omega_3} (Z^i_j)^2 > \frac{12}{(n_g-2)(n_g-3)} < 1.$$  

From (23), it is known that $A_{11}$ is the principal component to be computed when computing $A_{11}$. We can get similar results when calculating $C_{s,t}$. Above all, the simplified test statistic can be represented as

$$V_1 = \sum_{\{i, j\} \in \Omega_1} \frac{2}{Q_n^2} \left((Z^i_j)^2 + (Z^i_j)^2\right) - \frac{2}{n_g^2} (Y_{ij})^2.$$  

Let $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$. It is noted that this kind of approximate computation will reduce the computation from $O(n^4_g)$ to $O(n^2_g)$. The price paid for such an operation is that the simplified statistic in (24) is no longer unbiased.

### APPENDIX C

### ADDITIONAL CASE STUDY RESULTS

The signals were generated in load of P-V node $\rho = 19$ and $\rho = 1044$ for the case of IEEE 30-bus and the Polish 2383-bus system, respectively. Other experimental conditions were the same as the tests for the IEEE 118-bus system. The experiment results are shown in Fig. 7, Fig. 8, Fig. 9, and Fig. 10.

---

**Fig. 7:** Parameter learning of IEEE 30-bus system

---

### REFERENCES

[1] V. Terzija, G. Valverde, D. Cai, P. Regulski, V. Madani, J. Fitch, S. Skok, M. M. Begovic, and A. Phadke, “Wide-area monitoring, protection, and control of future electric power networks,” Proceedings of the IEEE, vol. 99, no. 1, pp. 80–93, 2011.

[2] R. C. Qiu and P. Antonik, Smart Grid and Big Data: Theory and Practice. Hoboken, NJ, USA: John Wiley & Sons, 2015.

[3] C. Lu, B. Shi, X. Wu, and H. Sun, “Advancing china’s smart grid: Phasor measurement units in a wide-area management system,” IEEE Power and Energy Magazine, vol. 13, no. 5, pp. 60–71, 2015.

[4] S. Nuthalapati and A. G. Phadke, “Managing the grid: Using synchrophasor technology [guest editorial],” IEEE Power and Energy Magazine, vol. 13, no. 5, pp. 10–12, 2015.

[5] S. Chakrabarti, E. Kyriakides, T. Bi, D. Cai, and V. Terzija, “Measurements get together,” IEEE Power and Energy Magazine, vol. 7, no. 1, pp. 41–49, 2009.

[6] J. De La Ree, V. Centeno, J. S. Thorp, and A. G. Phadke, “Synchronized phasor measurement applications in power systems,” IEEE Transactions on Smart Grid, vol. 1, no. 1, pp. 20–27, 2010.

[7] E. Pan, D. Wang, and Z. Han, “Analyzing big smart metering data towards differentiated user services: A
Fig. 8: Determination of most sensitive bus for IEEE 30-bus system. The system were effected by Type I signal with GSN, Type I signal with GMN, Type II signal with GSN, Type II signal with GNN, Type III signal with GSN and Type III signal with GMN, respectively.

Fig. 9: Parameter learning of the Polish 2383-bus system

(a) Parameter learning with GSN. (b) Parameter learning with GMN.

Fig. 10: Determination of most sensitive bus for the Polish 2383-bus system. (The system were effected by Type I signal with GSN, Type I signal with GMN, Type II signal with GSN, Type II signal with GNN, Type III signal with GSN and Type III signal with GMN, respectively.)

sublinear approach,” IEEE Transactions on Big Data, vol. 2, pp. 249–261, 2016.

[8] P.-A. Lof, G. Anderson, and D. Hill, “Voltage stability indices for stressed power systems,” IEEE Transactions on Power Systems, vol. 8, no. 1, pp. 326–335, 1993.

[9] S. G. Ghiocel and J. H. Chow, “A power flow method using a new bus type for computing steady-state voltage stability margins,” IEEE Transactions on Power Systems, vol. 29, no. 2, pp. 958–965, 2014.

[10] I. R. Pordanjani, Y. Wang, and W. Xu, “Identification of critical components for voltage stability assessment using channel components transform,” IEEE Transactions on Smart Grid, vol. 4, no. 2, pp. 1122–1132, 2013.

[11] D. H. A. Lee, “Voltage stability assessment using equivalent nodal analysis,” IEEE Transactions on Power Systems, vol. 31, no. 1, pp. 454–463, 2016.

[12] X. Xu, X. He, Q. Ai, and R. C. Qiu, “A correlation analysis method for power systems based on random matrix theory,” 2015, DOI: 10.1109/TSG.2015.2508506.

[13] L. Xie, Y. Chen, and P. Kumar, “Dimensionality reduction of synchrophasor data for early event detection: Lin-
earized analysis,” IEEE Transactions on Power Systems, vol. 29, no. 6, pp. 2784–2794, 2014.

[14] J. M. Lim and C. L. DeMarco, “Svd-based voltage stability assessment from phasor measurement unit data,” IEEE Transactions on Power Systems, vol. 31, no. 4, pp. 2557–2565, 2016.

[15] G. Ghanavati, P. D. Hines, and T. I. Lakoba, “Identifying useful statistical indicators of proximity to instability in stochastic power systems,” IEEE Transactions on Power Systems, vol. 31, no. 2, pp. 1360–1368, 2016.

[16] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, “A big data architecture design for smart grids based on random matrix theory,” 2015, DOI: 10.1109/TSG.2015.2445828.

[17] X. He, R. C. Qiu, Q. Ai, L. Chu, X. Xu, and Z. Ling, “Designing for situation awareness of future power grids: An indicator system based on linear eigenvalue statistics of large random matrices,” DOI: 10.1109/ACCESS.2016.2581838.

[18] Z. Bai and H. Saranadasa, “Effect of high dimension: by an example of a two sample problem,” Statistica Sinica, pp. 311–329, 1996.

[19] A. K. Gupta and J. Xu, “On some tests of the covariance matrix under general conditions,” Annals of the Institute of Statistical Mathematics, vol. 58, no. 1, pp. 101–114, 2006.

[20] Z. Bai, D. Jiang, J.-F. Yao, and S. Zheng, “Corrections to lrt on large-dimensional covariance matrix by rmt,” The Annals of Statistics, pp. 3822–3840, 2009.

[21] O. Ledoit and M. Wolf, “Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size,” Annals of Statistics, pp. 1081–1102, 2002.

[22] S. X. Chen and Y.-L. Qin, “A two-sample test for high-dimensional data with applications to gene-set testing,” The Annals of Statistics, vol. 38, no. 2, pp. 808–835, 2010.

[23] S. X. Chen, L.-X. Zhang, and P.-S. Zhong, “Tests for high-dimensional covariance matrices,” Journal of the American Statistical Association, 2012.

[24] A. Lee, “U-statistics,” Theory and Practice,” Marcel Dekker, New York, 1990.

[25] F. Milano and R. Zárate-Miñano, “A systematic method to model power systems as stochastic differential algebraic equations,” IEEE Transactions on Power Systems, vol. 28, no. 4, pp. 4537–4544, 2013.

[26] M. H. Bollen, Understanding power quality problems. IEEE press New York, 2000, vol. 3.

[27] F. Dorfler and F. Bullo, “Kron reduction of graphs with applications to electrical networks,” IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 60, no. 1, pp. 150–163, 2013.

[28] P. Kundur, J. Paserba, V. Ajjarapu, G. Andersson, A. Bose, C. Canizares, N. Hatziargyriou, D. Hill, A. Stankovic, and C. Taylor, “Definition and classification of power system stability ieee/cigre joint task force on stability terms and definitions,” IEEE transactions on Power Systems, vol. 19, no. 3, pp. 1387–1401, 2004.

[29] T. W. Anderson, “An introduction to multivariate statistical analysis, second edition, wiley, new york,” American Mathematical Monthly, vol. 66, no. 5, pp. 981–982, 1959.

[30] D. R. Cox and D. V. Hinkley, Theoretical statistics. CHAPMAN and HALL, 1974.

[31] R. Kavasseri and S. K. Srinivasan, “Joint placement of phasor and power flow measurements for observability of power systems,” IEEE Transactions on Power Systems, vol. 26, no. 4, pp. 1929–1936, 2011.

[32] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, “Matpower: Steady-state operations, planning, and analysis tools for power systems research and education,” IEEE Transactions on power systems, vol. 26, no. 1, pp. 12–19, 2011.

[33] R. D. Zimmerman and C. E. Murillo-Sanchez, “Matpower 5.1-user’s manual,” Power Systems Engineering Research Center (PSERC), 2015.

[34] M. HALL, “Permutations and combinations,” Combinatorial Theory, Second Edition, pp. 1–7, 1983.

Lei Chu has been pursuing the Ph.D degree at Shanghai Jiaotong University, since 2015. He is also a research assistant in the Big Data Engineering Technology and Research Center, Shanghai. He has published two book chapters, several journal papers and patents. His research interests are in the theoretical and algorithmic studies in signal processing, statistical learning for high dimensional data, random matrix theory, deep learning, as well as their applications in wireless communications, bioengineering and smart grid.
Robert Qiu (IEEE S’93-M’96-SM’01-FM14) received the Ph.D. degree in electrical engineering from New York University (former Polytechnic University, Brooklyn, NY). He is currently a Professor in the Department of Electrical and Computer Engineering, Center for Manufacturing Research, Tennessee Technological University, Cookeville, Tennessee, where he started as an Associate Professor in 2003 before he became a Professor in 2008. He has also been with the Department of Electrical Engineering, Research Center for Big Data Engineering and Technologies, State Energy Smart Grid R&D Center, Shanghai Jiaotong University since 2015. His current interest is in wireless communication and networking, machine learning and the Smart Grid technologies. He was Founder-CEO and President of Wiscom Technologies, Inc., manufacturing and marketing WCDMA chipsets. Wiscom was sold to Intel in 2003. Prior to Wiscom, he worked for GTE Labs, Inc. (now Verizon), Waltham, MA, and Bell Labs, Lucent, Whippany, NJ. He has worked in wireless communications and network, machine learning, Smart Grid, digital signal processing, EM scattering, composite absorbing materials, RF microelectronics, UWB, underwater acoustics, and fiber optics. He holds over 6 patents and authored over 70 journal papers/book chapters and 90 conference papers. He has 15 contributions to 3GPP and IEEE standards bodies. In 1998 he developed the first three courses on 3G for Bell Labs researchers. He served as an adjunct professor in Polytechnic University, Brooklyn, New York. Dr. Qiu serves as Associate Editor, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and other international journals. He is a co-author of Cognitive Radio Communication and Networking: Principles and Practice (John Wiley), 2012 and Cognitive Networked Sensing: A Big Data Way (Springer), 2013, and the author of Big Data and Smart Grid (John Wiley), 2015. He is a Guest Book Editor for Ultra-Wideband (UWB) Wireless Communications (New York: Wiley, 2005), and three special issues on UWB including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and IEEE TRANSACTION ON SMART GRID. He serves as a Member of TPC for GLOBECOM, ICC, WCNC, MILCOM, ICUWB, etc. In addition, he served on the advisory board of the New Jersey Center for Wireless Telecommunications (NJCWT). He is included in Marquis Who’s Who in America.

Xing He received the bachelors and masters degree in electrical engineering from Southeast University, Nanjing and Shanghai Jiaotong University, Shanghai, in 2008 and 2012, respectively. Currently he is pursuing his Ph.D. studies at School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai, China. His research interests include big data modeling and anomaly detection.

Zenan Ling received the bachelors degree in Department of Mathematics from Nanjing University in 2015. He has been pursuing the Ph.D degree at Shanghai Jiaotong University, since 2015. He is also a research assistant in the Big Data Engineering Technology and Research Center, Shanghai. His research interests are in the machine learning, random matrix and free probability as well as their application in smart grid, stock market and image processing.

Yadong Liu was born in Wuhan, China. He received his M.D and Ph.D. of electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2008 and 2012 respectively. Currently, he is a lecturer in school of electronic information and electrical engineering at SJTU. His research interests include detecting fault location and diagnosis of power equipment.