A Synchrophasor Data Compression Technique With Iteration-Enhanced Phasor Principal Component Analysis

Fang Zhang, *Senior Member, IEEE*, Xiaojun Wang, *Member, IEEE*, Ying Yan, *Student Member, IEEE*, Jinghan He, *Fellow, IEEE*, Wenzhong Gao, *Fellow, IEEE*, and Gang Chen, *Member, IEEE*

***Abstract*—The phasor data were compressed as separated amplitudes and phases in previous synchrophasor data compression techniques. To utilize the spatial correlation and temporal continuity of synchrophasors for data compression, a phasor principal component analysis (PPCA) in the field of complex numbers is proposed to compress synchrophasors as a whole in this article. Then, an iterative phasor principal components selection method is proposed to achieve PPCA and ensure the accuracy of reconstructed data since the existing eigenvalue-based criteria are not suitable for data compressions. Moreover, the proposed PPCA is enhanced by an iteration-based process to reduce the computation of PPCA. Actual PMU data measured under both a low-frequency oscillation incident and a two-phase short circuit incident conditions are used to verify the performance of PPCA compared with a recent PCA-based compression method. The results demonstrate that PPCA achieves higher compression ratios with better accuracy of reconstructed data, significantly reduced computation, and better real-time performance under both conditions.**

***Index Terms*—Complex numbers, data compression, iterative methods, phasor measurement unit, principal component analysis, wide-area measurement system.**

I. Introduction

PHASOR Measurement Units (PMU) and Wide-area Measurement Systems (WAMS) provide synchronized phasor measurements with considerably higher reporting rates. The massive amount of PMU measurement data is a severe factor that limits further applications of synchronous measurements. Since most of the PMU data are phasor data, also known as synchrophasors, the compression of phasor data will directly affect the compression ratio and the accuracy of reconstructed data, which are two leading indicators for the performance of data compression. Therefore, synchrophasor data compression techniques are needed to achieve higher efficiency of PMU data storage and communication [1]–[8].

Data compression techniques mainly include the lossless and lossy data compressions. Recent lossless data compression techniques are studied in [7]–[10], and these techniques are usually improved from data compression for general purpose. Lossless data compressions achieve absolutely accurate reconstructed data at the expense of much lower compression ratios; thus, they are suitable for the situations that require absolutely accurate reconstructed data.

Meanwhile, lossy data compression may achieve much higher compression ratios at the expense of limited error of reconstructed data. Three main types of lossy data compression methods are studied previously, including the feature extraction [4], [5], [11]–[18], trend extraction [1], [2] and parametric coding [3], [19]. Feature extraction is the most studied type currently, especially for the wavelet-based data compression techniques [11]–[15], [20]. As summarized in [3], wavelet-based techniques are multi-resolution decompositions and studied as three main branches, including the Discrete Wavelet Transform (DWT) [11]–[13], Wavelet Packet Transform (WPT) [15], and Embedded Zero-tree Wavelet (EZW) [14]. Principal component analysis (PCA) is another widely-used spatial feature extraction [5], [6], [21]–[23]. Moreover, matrix decomposition [4], [16] and convolutional autoencoder [18] can also be used for feature extraction . Trend extraction techniques pick much fewer data points from the raw data that can represent the trend of raw data to achieve data compression, e.g., the swing door trending (SDT) [1], [2]. Parametric coding techniques first model the raw data with a few parameters and then compress the raw data by obtaining these parameters. For example, the raw data are modeled by the damped sinusoids in [3], or the spectral shape [19], [20]. However, the compression objects of all these methods are real numbers. In particular, the studies in [1]–[8], [22] specifically proposed PMU data compression techniques while the voltage and current phasors are still compressed as separated amplitudes and phases in real numbers.

A serious disadvantage of these studies is that the amplitudes and phases of a single voltage or current phasor are treated as two independent data sequence and processed separately. This disadvantage leads to two impacts. The first impact

Manuscript received December 18, 2019; revised May 15, 2020 and October 15, 2020; accepted December 13, 2020. Date of publication December 22, 2020; date of current version April 21, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 52077004, and in part by the Fundamental Research Funds for the Central Universities under Grant 2019RC006. Paper no. TSG-01886-2019. *(Corresponding authors: Fang Zhang; Xiaojun Wang.)*

Fang Zhang, Xiaojun Wang, Ying Yan, and Jinghan He are with the School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: thu.zhangfang@gmail.com; xjwang1@bjtu.edu.cn).

Wenzhong Gao is with the Department of Electrical and Computer Engineering, University of Denver, Denver, CO 80208 USA.

Gang Chen is with the Department of Power System, State Grid Sichuan Electric Power Research Institute, Chengdu 610041, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2020.3046666.

Digital Object Identifier 10.1109/TSG.2020.3046666

is that the one-to-one correspondence between the amplitudes and phases of a phasor is ignored, and thus this valuable information is wasted. Another impact is that, although the accuracy of reconstructed amplitudes and phases can be guaranteed within a certain range in the existing compression methods, the errors of the reconstructed phasors composed of the reconstructed amplitudes and phases are difficult to be guaranteed, because the broken correspondence will cause additional errors for sure, as concluded in Section II-A. Therefore, data compression methods in the field of complex numbers are needed to simultaneously compress the amplitudes and phases of a phasor.

By an orthogonal transformation, PCA-based techniques can compress raw data into a set of linearly uncorrelated variables called principal components, thus minimizing the spatial redundancy of raw data. The reconstructed data of compressions are the linear combination of principal components (PC), and the linear coefficients are the eigenvectors of the covariance matrix of the raw data [5], [6], [23]. In [5], an efficient PCA-based method that compresses the amplitudes and phases separately, denoted by EPCA here, is proposed for synchrophasors and discussed the characteristics of PCA. EPCA is used as a comparison in this article. However, three essential drawbacks of the previous PCA-based techniques for synchrophasor data compression still exist though many related studies are proposed. First, PCA-based data compressions in real-numbers cannot deal with phasors as complex numbers. Second, a particular method for the selection of phasor principal components (PPC) is needed to keep the relevant information after compression. Two criteria to determine the principal components (PC) are commonly used in conventional PCAs. One is the normalized cumulative variance [5], [6], [17] and the other is the Guttman lower bound criterion [24], [25]. Both of the criteria determine the PC statistically according to the eigenvalues of the covariance matrix. When used for data compression, these criteria will no longer work well. These criteria cannot maintain a sufficient accuracy of the reconstructed data, because the eigenvalues of the covariance matrix are determined by the characteristics of the raw data and do not correspond to the accuracy of the reconstructed data. Third, a large data window of samples is required to extract the common features in the raw data accurately and efficiently. Moreover, the large data window causes the computation burden and the poor real-time performance of PCA-based data compression techniques [5]. To take advantage of the one-to-one correspondence between the amplitudes and phases of a phasor, the previous study [17] of this article attempted to perform the complex PCA [24], [26] for synchrophasor data compression as the prototype of Phasor PCA (PPCA); however, the prototype of PPCA is not practical enough since it still suffers from the above second and third drawbacks.

On the basis of the previous work [17], the PPCA is enhanced by solving these essential issues in the selection of PPCs and the large amount of computation for practical applications of synchrophasor data compression in this article. The features of the proposed iteration-enhanced PPCA synchrophasor data compression are in three aspects,

1) The PPCA synchrophasor data compression is improved from PCA in the field of complex numbers so that the correlation between amplitudes and phases of synchrophasors is utilized to achieve better compression performance.
2) An iterative PPC selection method is proposed. The accuracy of reconstructed data is taken as the index for the iteration, and it is thus guaranteed directly for any conditions in power systems. Also, this iterative selection method will not increase the computation complexity.
3) The iteration-based process for PPCA reduces the computation of PPCA significantly. By taking advantage of the temporal continuity of synchrophasors, the spatial correlation between the PPCs and the raw data is stable unless disturbances occur. That is, PPCA will not be calculated until disturbances occur and the spatial correlation changes. Thus, PPCA achieves better real-time performance with overlapping data windows.

Finally, the proposed iteration-enhanced PPCA technique is verified with actual PMU data under both a low-frequency oscillation incident and a two-phase short circuit incident conditions, respectively. Also, PPCA is compared with a recent EPCA method [5] that compresses the amplitudes and phases separately. The comparison results demonstrate that PPCA can achieve the guaranteed higher accuracy of reconstructed data with higher compression ratios, significantly reduced computation, and better real-time performance.

This article is organized as follows, PPCA method is proposed in Section II; the iterative PPC selection method is proposed in Section III; the iteration-based process for PPCA is proposed in Section IV; the validation of the proposed method is illustrated with actual PMU data in Section V; the conclusions are given in Section VI.

II. Phasor Principal Component Analysis (PPCA)

The basic idea of the proposed PPCA method is to extract the physically defined spatial and temporal features from the raw phasor data matrix in the complex domain, and then minimize the spatial redundancy among the raw data on the phasor principal components standing for different temporal features. Moreover, PPCA is enhanced by the two-level iterative algorithms for better compression performance and less computation, as shown in Fig. 1.

A. Basic Idea of PPCA

In power systems, e.g., Fig. 1, the raw phasor data matrix $\mathbf{D_r}$ is measured by PMUs as,

$$\mathbf{D_r}=\begin{bmatrix} \dot{X}_1(1) & \cdots & \dot{X}_j(1) & \cdots & \dot{X}_N(1) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \dot{X}_1(i) & \cdots & \dot{X}_j(i) & \cdots & \dot{X}_N(i) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \dot{X}_1(M) & \cdots & \dot{X}_j(M) & \cdots & \dot{X}_N(M) \end{bmatrix}, \begin{matrix} 1 \le i \le M \\ 1 \le j \le N \\ \mathbf{D_r} \in \mathbb{C}_{M\times N} \end{matrix}. \tag{1}$$

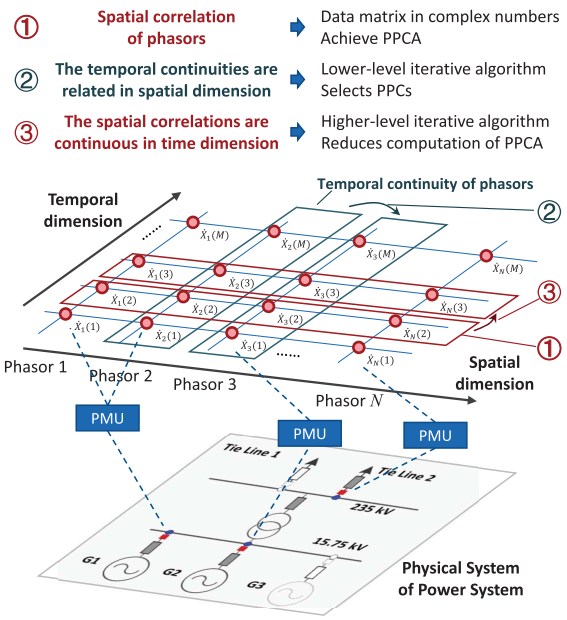


Fig. 1. Basic idea of PPCA: Spatial correlation and temporal continuity of synchrophasors.

A single phasor is calculated from instantaneous values and expressed as the amplitude and phase with simultaneous correspondence; hence, the physical system variations directly cause the continuous change in instantaneous values and affect the continuous coupling change in the amplitude and phase of the phasor simultaneously. Thus, a series of a single phasor, e.g., $\dot{X}_j(1), \ldots, \dot{X}_j(M)$, have the feature of temporal continuity.

Meanwhile, for different phasors, the relationship between the voltage and current phasors is directly determined by the physical system of the power system, i.e., the electrical network, according to the Kirchhoff Voltage and Current Laws (KVL & KCL) as $\dot{\mathbf{I}} = \mathbf{Y}\dot{\mathbf{U}}$, where $\dot{\mathbf{I}}$, $\dot{\mathbf{U}}$, and $\mathbf{Y}$ are current phasors, voltage phasors, and nodal admittance matrix, respectively. Therefore, phasors at the same time of different nodes, e.g., $\dot{X}_1(i), \ldots, \dot{X}_N(i)$, have the feature of spatial correlation defined by the Kirchhoff laws. For example, the PMU data of the actual substation in Fig. 9 of Section V-B include 40 voltage phasors and 40 current phasors covering 10 different measuring points. Only six phasors are independent of each other in a steady state; while, the others can be calculated from them. In addition, although $\dot{\mathbf{I}} = \mathbf{Y}\dot{\mathbf{U}}$ in phasor form may not strictly exist when disturbances occur, the instantaneous values of voltage and current are always determined by the Kirchhoff laws. Thus, the spatial correlation of phasors still exists during disturbances, but disturbances will complicate it, and more compressed data will be needed to save disturbance information.

In previous studies [5], [6], [21], the spatial correlation between the amplitudes or the phases of different phasors is utilized for data compression; however, this spatial correlation is simplified and weakened from the spatial correlation of phasor $\dot{\mathbf{I}} = \mathbf{Y}\dot{\mathbf{U}}$, as the result of the one-to-one correspondence between the amplitude and phase of a phasor. Consequently, some physically defined valuable spatial correlations are lost when compressing amplitudes and phases separately, the features of the raw data are thus weakened, resulting in reduced compression performance. This fact is verified by the illustrations in Section V.

Thus, compressing phasors as complex numbers will facilitate the extraction of common features and can thus achieve better compression performance by utilizing the one-to-one correspondence between the amplitudes and phases of a phasor. Three characteristics of the proposed PPCA method are shown in Fig. 1 and as follows:

1) PPCA process PCA for the phasor data matrix $\mathbf{D_r}$ in complex numbers, and thus utilize the spatial correlation features in the rows of $\mathbf{D_r}$ and the temporal continuity features in the columns of $\mathbf{D_r}$.
2) The trends of different phasors in the temporal dimension are similar and related to each other, particularly the temporal continuities of phasors are related in the spatial dimension. Thus, a lower-level iterative algorithm is proposed to selects PPCs iteratively to control the accuracy of reconstructed phasors, since only a few iterations will keep enough PPCs unless during disturbances.
3) The spatial correlations of different phasors at adjacent time change rarely, particularly the spatial correlations of phasors are continuous in the temporal dimension. Thus, a higher-level iterative algorithm is proposed to reduce the computations of PPCA, since the spatial correlation requires complete recalculation only when disturbances change it.

B. Process of PPCA

Similar to the conventional PCA [5], [6] and the complex PCA [24], [26], PPCA compresses the information contained in the covariance matrix $\mathbf{C}$ into a relatively few complex eigenvectors with elements $\mathbf{u_i}$ and PPCs $\mathbf{p_i}$ in complex numbers where $i = 1, \ldots, N'$, N' is the number of PPCs for PPCA and $N' < N$ generally [17]. The key differences of PPCA from the complex PCA are the normalization of data matrix and the selection of PPCs. The flow chart of PPCA with the amount of computation for each step is shown in Fig. 2.

1) Formation of Data Matrix: The first step of PPCA is to form the data matrix. As explained in Section II-A, the data matrix for PPCA consists of the voltage phasors and current phasors of a single PMU or PDC, as,

$$\mathbf{D_r} = \left(\dot{\mathbf{U}}_1, \ldots, \dot{\mathbf{U}}_{\mathbf{N_V}}, \dot{\mathbf{I}}_1, \ldots, \dot{\mathbf{I}}_{\mathbf{N_I}}\right), \tag{2}$$

where $\mathbf{D_r} \in \mathbb{C}_{M\times(N_V+N_I)}$, M is the number of data samples for each phasor in a duration of T seconds with a sampling rate of F_s samples per second and thus $M = T \times F_s$; N_V and N_I are the number of voltage and current phasors of all PMUs, respectively. The following contents will not distinguish between voltage and current phasors unless otherwise stated. Phasors are uniformly denoted by $\dot{X}$. Let N denote the number of phasors, i.e., $N = N_V + N_I$, thus $\mathbf{D_r} \in \mathbb{C}_{M\times N}$ with $M \gg N$ as,

$$\mathbf{D_r} = \left(\dot{\mathbf{X}}_1, \ldots, \dot{\mathbf{X}}_\mathbf{j}, \ldots, \dot{\mathbf{X}}_\mathbf{N}\right), 1 \le j \le N, \mathbf{D_r} \in \mathbb{C}_{M\times N}, \tag{3a}$$

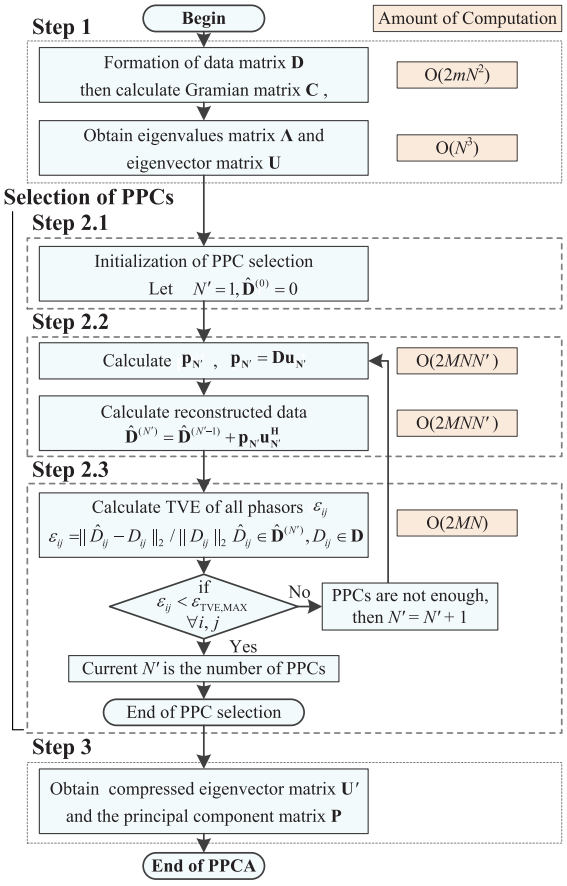


Fig. 2. Flow chart of PPCA with iterative phasor principal components selection method.

$$\dot{\mathbf{X}}_\mathbf{j} = \left(V_{1j}\angle\alpha_{1j}, \ldots, V_{ij}\angle\alpha_{ij}, \ldots, V_{Mj}\angle\alpha_{Mj}\right)^{\mathrm{T}}, 1 \le i \le M, \tag{3b}$$

where V and α stand for the amplitude and phase of $\dot{X}$.

2) Normalization of Data Matrix: Before the further steps of PPCA, the data matrix of PPCA is normalized first so that the variances of different synchrophasors are on the same order for better extraction of common features. In PPCA, the synchrophasor samples are normalized to the normalized phasors with the mean amplitude of 1 without changing the phases. Thus, the normalization matrix of $\mathbf{D_r}$ is $\mathbf{A_N}$ that,

$$\mathbf{A_N} = \mathrm{diag}\left(\frac{\sum_{i=1}^{M} V_{i1}}{M}, \ldots, \frac{\sum_{i=1}^{M} V_{ij}}{M}, \ldots, \frac{\sum_{i=1}^{M} V_{iN}}{M}\right), \quad 1 \le j \le N. \tag{4}$$

Then the normalized data matrix of $\mathbf{D_r}$, denoted by $\mathbf{D}$, is that,

$$\mathbf{D} = \mathbf{D_r}\mathbf{A_N}^{-1}, \mathbf{D} \in \mathbb{C}_{M\times N}. \tag{5}$$

Note that the normalization of the data matrix in PPCA is different from that in conventional PCAs where the samples are normalized with a mean of 0 and a standard deviation of 1 as a result of the statistical criteria for PPC selection [5], [6], [26].

3) Selection of Principal Components: The covariance matrix $\mathbf{C}$ of complex matrix $\mathbf{D}$ [26] is calculated as,

$$\mathbf{C} = \frac{1}{M-1}\mathbf{D}^{\mathrm{H}}\mathbf{D}, \mathbf{C} \in \mathbb{C}_{N\times N}, \tag{6}$$

where $\mathbf{D}^{\mathrm{H}}$ is the conjugate transpose or Hermitian transpose of $\mathbf{D}$. Then the covariance matrix $\mathbf{C}$ is a Hermitian matrix and a semi-positive definite matrix. When using overlapping data windows for PPCA as in Section IV, only the elements of $\mathbf{C}$ that relate to the new data are calculated; thus, the amount of computation for $\mathbf{C}$ is $\mathrm{O}(2mN^2)$ where m is the number of new data samples for each phasor and O(*) stands for the Big O notation for the computation complexity.

As $\mathbf{C}$ is a Hermitian matrix, the eigenvalues λ_i of $\mathbf{C}$ are real numbers and the eigenvector matrix $\mathbf{U}$ is a Unitary matrix that,

$$\begin{aligned}&\mathbf{\Lambda} = \mathbf{U}^{\mathrm{H}}\mathbf{C}\mathbf{U},\\ &\mathbf{U} = [\mathbf{u_1}, \ldots, \mathbf{u_i}, \ldots, \mathbf{u_N}], \mathbf{u_i} \in \mathbb{C}_{N\times 1},\\ &\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_i, \ldots, \lambda_N), \lambda_1 \ge \cdots \ge \lambda_N \ge 0, \lambda_i \in \mathbb{R}.\end{aligned} \tag{7}$$

The method to select PPCs and the determine number of PPC N' for PPCA is proposed in Section III. Then the compressed eigenvector matrix $\mathbf{U}'$ and the principal component matrix $\mathbf{P}$ can be obtained as,

$$\mathbf{U}' = [\mathbf{u_1}, \ldots, \mathbf{u_{N'}}], \mathbf{U}' \in \mathbb{C}_{N\times N'} \tag{8a}$$

$$\mathbf{P} = \left[\mathbf{p_1}, \ldots, \mathbf{p_{N'}}\right] = \mathbf{D}\mathbf{U}', \mathbf{P} \in \mathbb{C}_{M\times N'}. \tag{8b}$$

The complex matrix $\mathbf{U'}_{N\times N'}$, $\mathbf{P}_{M\times N'}$ and $\mathbf{A_N}$ are the compressed data. The compression ratio λ_{CR} defined as (9a) in [1], [5], etc. can be calculated as (9b) due to $N \ll M$ generally,

$$\lambda_{\mathrm{CR}} = \frac{\text{the number of } \textit{Raw Data}}{\text{the number of } \textit{Compressed Data}} \tag{9a}$$

$$= \frac{M\times N}{N\times N' + M\times N' + N} \approx \frac{N}{N'}. \tag{9b}$$

4) Reconstruction of Compressed Data: The reconstructed phasor data are the estimates of compressed phasors. According to the mathematical explanation of compressed data of PPCA as (8), the reconstructed phasor data of $\mathbf{D}$, denoted by $\hat{\mathbf{D}}$ where the mark "ˆ" stands for "estimate", is calculated as,

$$\hat{\mathbf{D}} = \mathbf{P}\mathbf{U'}^{\mathrm{H}}, \hat{\mathbf{D}} \in \mathbb{C}_{M\times N}. \tag{10}$$

Then the final reconstructed data, denoted by $\hat{\mathbf{D}}_r$, is calculated by the denormalization of $\hat{\mathbf{D}}$ as

$$\hat{\mathbf{D}}_r = \hat{\mathbf{D}}\mathbf{A_N}, \hat{\mathbf{D}}_r \in \mathbb{C}_{M\times N}, \tag{11}$$

and $\hat{\mathbf{D}}_r$ is the estimate of the raw phasors in $\mathbf{D_r}$ according to the compressed data $\mathbf{U'}_{N\times N'}$, $\mathbf{P}_{M\times N'}$, and $\mathbf{A_N}$. Equation (10) indicates that the raw phasors $\mathbf{D}$ are compressed as the linear combination of the columns of $\mathbf{P}_{M\times N'}$, i.e., the PPCs, and the coefficients are the conjugations of each row of $\mathbf{U'}_{N\times N'}$. That is, $\mathbf{p_i}$ in $\mathbf{P}$ stands for the features of temporal continuity of the phasors; meanwhile, $\mathbf{u_i}$ in $\mathbf{U}'$ stands for the features of spatial correlation between the phasors from different nodes, as illustrated in Fig. 1.

The *total vector error* (TVE) ε_{TVE} defined by IEEE standard [27] is used in this article to evaluate the error between the estimated and the raw phasors as (12),

$$\varepsilon_{\text{TVE}} = \frac{\|\dot{\hat{X}} - \dot{X}\|_2}{\|\dot{X}\|_2} \times 100\%, \tag{12}$$

where $\dot{\hat{X}}$ and $\dot{X}$ stand for a single data point of the corresponding reconstructed phasor and raw phasor, respectively.

III. Iterative Selection of Phasor Principal Components

An iterative PPC selection method is proposed for PPCA to enhance the compression performance of PPCA in this section. The choice of principal components N' remains one of the essential considerations in any PCA-based techniques so that all the relevant information is retained [5], [26].

In the existing studies, two criteria as the index of statistical significance were commonly used to determine the number of principal components N' in conventional PCAs [26]. The first criterion is the *normalized cumulative variance criterion* used in [5], [6], [17], where the normalized cumulative variance δ of N' principal components is $\delta = (\sum_{i=1}^{N'} \lambda_i)/(\sum_{i=1}^{N} \lambda_i)$ and the criterion is $N' = \min\{N'|\delta \geq \delta_{\text{criterion}}\}$. It is discussed and compared in Section V. The second criterion is the *Guttman lower bound criterion* [24]. It recommends retaining all of the principal components that contribute more total variance than the typical normalized time series, i.e., one unit of total variance that $N' = \max\{N'|\lambda_{N'} > 1\}$. This criterion is widely used in statistical science, for example, the well-known SPSS Statistics Software [25].

However, both of these criteria are based on the normalization of sample data so that the eigenvalues λ_i in $\mathbf{\Lambda}$ stands for the normalized variances of each principal component $\mathbf{p_i}$, respectively; where the normalized variances are the total square error of $\hat{\mathbf{D}}$ to $\mathbf{D}$ [5], [24], [26]. That is, these eigenvalue-based criteria determine the principal components by the statistical significance with the total square error of $\hat{\mathbf{D}}$ instead of the accuracy of each individual reconstructed data. Moreover, the statistical significance measured by the eigenvalues λ_i is affected by different disturbances; thresholds of the criteria are thus needed to be changed all the time for different disturbances to maintain a sufficient accuracy of reconstructed data. In [5], the disturbance detection is performed before EPCA to switch between different thresholds of normalized cumulative variance criterion for normal data and disturbance data, respectively. However, though the thresholds are finely differentiated for disturbance conditions, these statistical criteria may miss some particular features of certain individual data points; hence, unacceptably low accuracy of reconstructed data may occur for these individual data points. Thus, the circumstance that a dynamic snapshot [28] provided by PMU synchrophasors as in Fig. 3 suffers from low accuracy while the total square error is still satisfied may occur, as illustrated in Figs. 7 and 11 in Section V. Furthermore, due to the indirect control of accuracy, the eigenvalue-based criteria cannot work with the following iteration-enhanced PPCA in Section IV to reduce the amount of computation.

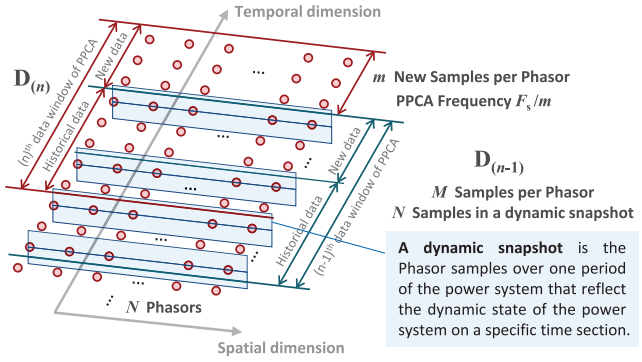


Fig. 3. Dynamic snapshot and overlapping data window of iteration-enhanced PPCA for better real-time performance.

The basic idea of the proposed iterative PPC selection method is to correct the estimate data matrix $\hat{\mathbf{D}}$ by adding PPCs with iterations till the TVEs of all elements of $\hat{\mathbf{D}}$ from $\mathbf{D}$ satisfy the accuracy criterion. The flow chart with the amount of computation of this method is shown in Fig. 2 as **Step 2**. The eigenvector matrix $\mathbf{U}$ and the eigenvalue matrix $\mathbf{\Lambda}$ are obtained previously by Step 1 as introduced in Section II-B with (6) to (7).

Step 2.1 The initial iteration condition is $N' = 1$ and $\hat{\mathbf{D}}^{(0)} = 0$.

Step 2.2 Calculate the N'th principal component $\mathbf{p_{N'}}$ as an iteration with,

$$\mathbf{p_{N'}} = \mathbf{D}\mathbf{u_{N'}}. \tag{13}$$

Then calculate the reconstructed data $\hat{\mathbf{D}}$ as another iteration with,

$$\hat{\mathbf{D}}^{(N')} = \hat{\mathbf{D}}^{(N'-1)} + \mathbf{p_{N'}}\mathbf{u}_{\mathbf{N'}}^{\mathbf{H}}. \tag{14}$$

Step 2.3 Calculate the TVEs of all phasors according to (12) that

$$\varepsilon_{ij} = \frac{||\hat{D}_{ij} - D_{ij}||_2}{||D_{ij}||_2} \times 100\%, \quad \begin{matrix} \hat{D}_{ij} \in \hat{\mathbf{D}}^{(N')} \\ D_{ij} \in \mathbf{D} \end{matrix}, \tag{15}$$

where $\hat{D}_{ij}$ and D_{ij} are the elements of $\hat{\mathbf{D}}^{(N')}$ and $\mathbf{D}$, respectively. If $\varepsilon_{ij} < \varepsilon_{\text{TVE,MAX}}\ \forall i, j$, current N' is selected as the number of PPCs; otherwise, the number of PPCs is not enough to satisfy the accuracy criterion, then N' should be greater as $N' = N' + 1$ and the iteration in Step 2.2 should be repeated. Consequently, the number of PPCs N' is $N' = \min\{N'|\varepsilon_{ij} < \varepsilon_{\text{TVE,MAX}}\ \forall i, j\}$.

The following two features of the proposed iterative PPC selection method make it suitable for synchrophasor data compression.

First, this method can control the accuracy of reconstructed data directly and thus ensure the accuracy of each reconstructed dynamic snapshot. When applied in statistics, PCA focuses on the extraction of statistical significance. Meanwhile, when applied to data compression, PCA should focus more on the consideration of compression ratios and ensure sufficient accuracy of reconstructed data in each dynamic snapshot instead. Thus, this method is more suitable

for data compression than the widely-used eigenvalue-based criteria.

Second, although this method implements iterations by repeating Steps 2.2 and 2.3, this will not increase the amount of computation of PPCA significantly. In the iteration process, when the determination if $\varepsilon_{ij} < \varepsilon_{\text{TVE,MAX}}\ \forall i,j$ is stopped, N' will increase to add more PPCs, and ε_{ij} will thus be monotonic-decreasing. Hence, after N' is increased, the determination if $\varepsilon_{ij} < \varepsilon_{\text{TVE,MAX}}\ \forall i,j$ should be started from where it is stopped rather than from the very beginning. Therefore, (15) will only be calculated $(MN + N' - 1)$ times and the total computation complexity of PPCA as in Fig. 2 is O(MNN'). Meanwhile, the total computation complexity of EPCA is also O(MNN') due to the calculation of PCs similar to (8b). Consequently, the total amount of PPCA computation will not increase even with the iterations. Also, the synchrophasors of a substation are three-phase symmetry in steady states and very similar to each other in most cases due to the close electrical distance; thus, the number of PPCs is rather small in most cases.

The accuracy criterion $\varepsilon_{\text{TVE,MAX}}$ can be selected as 0.5% to 1% according to actual needs, since the measurement accuracy of PMUs is recommended as 1% for steady states and 3% for dynamic states [27]. The TVE $\varepsilon_{\text{TVE,MAX}} = 1\%$ approximately equals to *normalized mean square error* (NMSE) as (19b) [1], [3] $\varepsilon_{\text{NMSE}} = 1 \times 10^{-4}$ or *normalized root mean square error* (NRMSE) [1] $\varepsilon_{\text{NRMSE}} = 1 \times 10^{-2}$, which are on the same level as other lossy data compressions.

IV. Iteration-Enhanced PPCA for Practical Applications

A significant drawback of PCA-based data compression methods is that a large data window of samples is required so that the common features in the raw data can be extracted accurately and efficiently according to the information within the raw data. Moreover, the large data window causes the computation burden and the poor real-time performance of PCA-based data compression techniques [5], [17]. To deal with this drawback, PPCA is enhanced by an iteration-based process with overlapping data windows to achieve significantly reduced computation and better real-time performance for practical applications.

The overlapping data window of iteration-enhanced PPCA is illustrated in Fig. 3. The data matrix $\mathbf{D}_{(n)}$ for nth PPCA contains $m \times N$ new data points and $(M-m) \times N$ historical data points from the $(n-1)$th PPCA data matrix $\mathbf{D}_{(n-1)}$. Thus, the iteration-enhanced PPCA data compression will be processed in the frequency with F_s/m (Hz). The smaller m is, the better real-time performance PPCA will achieve.

Compared to the case in statistics that the PCA objects are random variables, the synchrophasors in power systems are limited by the physical system; thus, the spatial correlations between different phasors typically have strong enough temporal continuities, as in Fig. 1. Consequently, the relationship between the raw data and PPCs described by the eigenvalue matrix tends to change with temporal continuities.

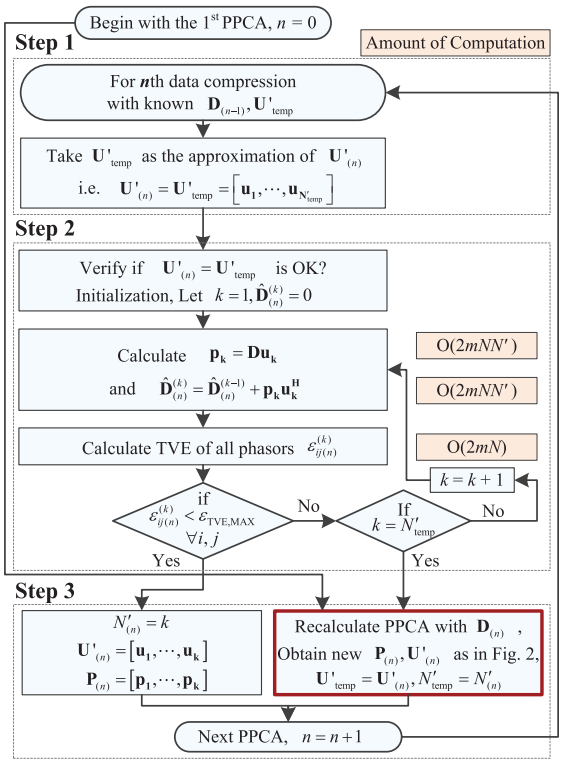


Fig. 4. Flow chart of iteration-enhanced PPCA for practical applications.

Therefore, the basic idea of iteration-enhanced PPCA is that the previously obtained eigenvector matrix can be directly used without being recalculated, when variations of power systems are not significant enough to change spatial correlations between different phasors, i.e., when the accuracy of reconstructed data is high enough. Moreover, the accuracy of reconstructed data can be used as the index to determine the recalculation and iterations. The process of the iteration-enhanced PPCA illustrated in Fig. 4 is as follows.

Step 0 Begin with the 1st PPCA data compression, obtain $\mathbf{P}_{(0)}$ and $\mathbf{U}'_{(0)}$, the number of PPCs $N'_{(0)}$, then set the intermediate variables $\mathbf{U}'_{\text{temp}}$ and N'_{temp} as $\mathbf{U}'_{\text{temp}} = \mathbf{U}'_{(0)}$ and $N'_{\text{temp}} = N'_{(0)}$, respectively.

Step 1 Start the nth PPCA data compression, $\mathbf{D}_{(n-1)}$ and $\mathbf{U}'_{\text{temp}}$ with N'_{temp} are known previously. Then, taking $\mathbf{U}'_{\text{temp}}$ as the approximation of $\mathbf{U}'_{(n)}$, i.e., $\mathbf{U}'_{(\mathbf{n})} = \mathbf{U}'_{\text{temp}} = [\mathbf{u_1}, \ldots, \mathbf{u}_{\mathbf{N}'_{\text{temp}}}]$.

Step 2 Determine if $\mathbf{U}'_{(\mathbf{n})} = \mathbf{U}'_{\text{temp}}$ meets the accuracy criteria of reconstructed data by increasing the number of PPCs in $\mathbf{U}'_{\text{temp}}$ one by one. The determination and the iterative criterion is similar to the iterative PPC selection in Section III. Calculate the kth principal component $\mathbf{p_k}$ as an iteration with,

$$\mathbf{p_k} = \mathbf{D}_{(\mathbf{n})}\mathbf{u_k}. \tag{16}$$

Then calculate the reconstructed data $\hat{\mathbf{D}}^{(k)}_{(n)}$ as another iteration with

$$\hat{\mathbf{D}}^{(k)}_{(n)} = \hat{\mathbf{D}}^{(k-1)}_{(n)} + \mathbf{p_k}\mathbf{u}^{\mathbf{H}}_{\mathbf{k}}. \tag{17}$$

Calculate the TVEs of each element in $\hat{\mathbf{D}}_{(n)}^{(k)}$ as,

$$\varepsilon_{ij(n)}^{(k)} = \frac{||\hat{D}_{ij(n)}^{(k)} - D_{ij}||_2}{||D_{ij}||_2} \times 100\%. \quad (18)$$

If $\varepsilon_{ij(n)}^{(k)} < \varepsilon_{\text{TVE,MAX}}$ holds for $\forall i, j$, current $\mathbf{U}'_{\text{temp}}$ with its first k columns is accurate enough as the approximation of $\mathbf{U}'_{(n)}$ that satisfies the accuracy criterion $\varepsilon_{\text{TVE,MAX}}$, and the recalculation of PPCA is not necessary. If not, more attempts will be made with iterations till $k = N'_{\text{temp}}$. When $k = N'_{\text{temp}}$ and the accuracy criterion is still not satisfied, $\mathbf{U}'_{\text{temp}}$ is not accurate enough as the approximation of $\mathbf{U}'_{(n)}$; then the recalculation of PPCA is needed. Note that the amount of computation here is about $O(mNN')$ which is different from that of (13), (14), and (15) in Section III, i.e., $O(MNN')$, because only the calculation that relates to the new data in $\mathbf{D}_{(n)}$ is needed for (16), (17), and (18).

Step 3 If the recalculation of PPCA is not necessary, then $N'_{(n)} = k$, $\mathbf{U}'_{(\mathbf{n})} = [\mathbf{u_1}, \ldots, \mathbf{u_k}]$, $\mathbf{P}_{(\mathbf{n})} = [\mathbf{p_1}, \ldots, \mathbf{p_k}]$, nth PPCA is finished. Otherwise, the complete PPCA for $\mathbf{D}(n)$ should be processed as Fig. 2 to obtain the new $\mathbf{U}'_{(n)}$ and $\mathbf{P}_{(n)}$ and refresh the intermediate variables $\mathbf{U}'_{\text{temp}}$ and N'_{temp} as $\mathbf{U}'_{\text{temp}} = \mathbf{U}'_{(n)}$ and $N'_{\text{temp}} = N'_{(n)}$, respectively. In the end, continue to the next PPCA with $n = n + 1$. Thus, the spatial correlation of phasors is stored as $\mathbf{U}'_{\text{temp}}$ and N'_{temp} until the next complete recalculation of PPCA is needed.

For better real-time performance, smaller size of the new data window m is needed. Since PPCA is performed in the frequency F_s/m (Hz), no matter how small m is, the total computation of iteration-enhanced PPCA in a certain duration T_d will be a constant as $O(T_d F_s NN')$; meanwhile, rare cases that require complete recalculations of PPCA will cause minimal impacts.

V. Verification With Actual Data

In this section, the proposed PPCA is verified with the raw data of synchrophasors measured by actual PMUs in both a low-frequency oscillation (LFO) incident and a two-phase short circuit incident under the conditions of both three-phase symmetric and asymmetrical faults, respectively. Moreover, both the results of PPCA, the prototype of PPCA proposed in [17], and the EPCA proposed in [5] are compared to demonstrate the performance and features of PPCA.

The accuracy criterion for iteration-enhanced PPCA is selected as $\varepsilon_{\text{TVE,MAX}} = 0.8\%$ in this section. The PPCA prototype uses the normalized cumulative variance criterion [17] with $N' = \min\{N'|\delta \geq 0.999\}$ and $N' = \min\{N'|\delta \geq 0.99999\}$ for Sections V-A and V-B, respectively. And the results with this condition are denoted by "PPCA-fixed". The principal components selection method for EPCA is the normalized cumulative variance criterion used in [5] with $N' = \min\{N'|\delta \geq 0.8\}$ for normal data and $N' = \min\{N'|\delta \geq 0.95\}$ for disturbance data, and the results with this condition are denoted by "EPCA-fixed". However, since the "EPCA-fixed" results are too poor, a dynamically changing normalized cumulative variance criterion $\delta_{\text{criterion}}$ that satisfies $\varepsilon_{\text{TVE,MAX}} = 1\%$ is applied for further comparison, and the results with this condition are denoted by "EPCA-dynamic". The compression ratio λ_{CR} defined by (9a) and the accuracy of reconstructed data are used for the evaluation of PPCA. Both the TVE defined in [27] as (12) and the *Normalized Root Mean Square Error* (NRMSE) (19) are used as the indicators of the accuracy of the reconstructed data. Noted that *Mean Square Error* (MSE) corresponds to the eigenvalue-based PPC selection criterion [26]; thus, the NRMSE can represent the statistical significance measured by the average errors.

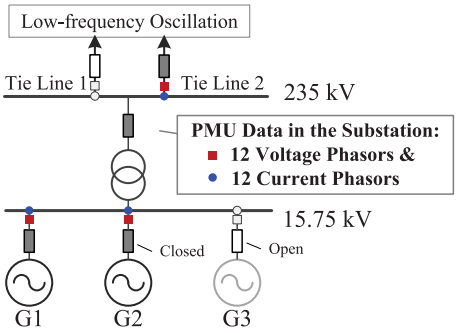


Fig. 5. Actual PMU data of the substation in the LFO incident.

$$\varepsilon_{\text{MSE}}\left(\hat{\mathbf{D}}, \mathbf{D}\right) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left\|\hat{D}_{ij} - D_{ij}\right\|_2^2}{MN}, \begin{matrix}\hat{D}_{ij} \in \hat{\mathbf{D}} \\ D_{ij} \in \mathbf{D}\end{matrix}, \quad (19a)$$

$$\varepsilon_{\text{NMSE}}\left(\hat{\mathbf{D}}, \mathbf{D}\right) = \frac{\varepsilon_{\text{MSE}}\left(\hat{\mathbf{D}}, \mathbf{D}\right)}{\varepsilon_{\text{MSE}}(\mathbf{D}, \mathbf{0})} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left\|\hat{D}_{ij} - D_{ij}\right\|_2^2}{\sum_{i=1}^{M}\sum_{j=1}^{N}\left\|D_{ij}\right\|_2^2}, \quad (19b)$$

$$\varepsilon_{\text{RMSE}} = \sqrt{\varepsilon_{\text{MSE}}}, \varepsilon_{\text{NRMSE}} = \sqrt{\varepsilon_{\text{NMSE}}}, \quad (19c)$$

The actual PMU data recorded in the actual power systems in this article can be found on the IEEE Dataport website as an open access resource [29].

A. PPCA in a Low-Frequency Oscillation Incident

In the LFO incident, the raw data are recorded in a substation of a hydropower plant as Fig. 5, which are the same as those in [1]. When the LFO incident occurred, G1, G2, and Tie Line 2 were in operation, whereas G3 and Tie Line 1 were in the outage state for maintenance. Each bus has the A-phase, B-phase, C-phase, and positive-sequence voltage and current synchrophasors. Thus, the raw data contain 12 voltage phasors and 12 current phasors. The total time interval for this demonstration is 250 s. The data window of PPCA contains $M = 1000$ samples of each phasor, and the duration T and the sampling rate F_s are 10 seconds and 100 Hz, respectively. Also, different sizes of overlapping data windows, i.e., conditions with different m, are used to illustrate the amount of computation with each method, respectively.

1) Raw Data: The raw data are presented and compared from the perspective of the synchrophasors as complex numbers and the amplitudes of synchrophasors as

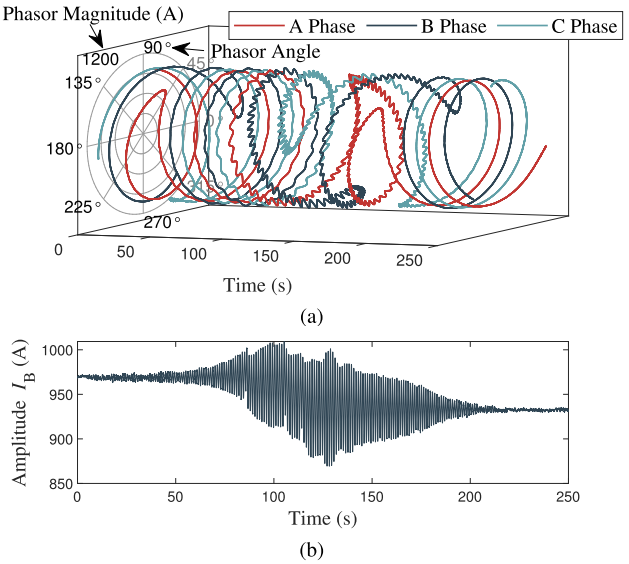


Fig. 6. Raw data of low-frequency incident. (a) Raw data of the three-phase current synchrophasors of G1 in cylindrical coordinates. (b) Raw data of amplitudes of the B-phase current synchrophasor of G1.

Figs. 6(a) and 6(b), respectively. The fluctuations of the synchrophasors as shown in Fig. 6(a) are more continuous with fewer distortions than those in the amplitudes of the same synchrophasors, as shown in Fig. 6(b). The cause is the fact that the amplitudes and phases of a phasor have a strong correlation and keep coupling with and affecting each other. Therefore, the phasor data compression in the field of complex numbers will take more advantages of phasor information for extracting features of disturbance data.

2) Compression Ratios: The compression ratios of both the proposed iteration-enhanced PPCA and EPCA over time are shown in Fig. 7(a). Note that each data point in Figs. 7 and 11 corresponds to a data window of data compression where the time of this data point indicates the middle of this data window. Before the LFO occurs (0 s to 80 s), the compression ratios of PPCA are kept at 12, i.e., $N' = 2$ and two PPCs are kept for voltage and current phasors, respectively. When the LFO occurs (80 s to 210 s), the compression ratios of PPCA reduce to 8, i.e., $N' = 3$ and one more PPC are kept to retain more disturbance information. After the LFO subsides (after 210 s), the compression ratios of PPCA increase to 12 again. Similarly, "PPCA-fixed" also has the same trend of compression ratio as PPCA but with higher compression ratios, even when "PPCA-fixed" suffers from bad accuracy as "EPCA-fixed" did in Fig. 7(b). On the other hand, the trend of the compression ratios of "EPCA-fixed" is significantly different from that of PPCA. Considering the increased $\delta_{\text{criterion}}$, the compression ratios of "EPCA-fixed" do not change when the LFO occurs; that is, "EPCA-fixed" does not keep more principal components during the LFO, which is opposite to PPCA. Furthermore, to find out the actual difference between PPCA and EPCA, the condition "EPCA-dynamic" with dynamically changing $\delta_{\text{criterion}}$ is applied. Though the accuracy of reconstructed data guaranteed by "EPCA-dynamic" is slightly lower

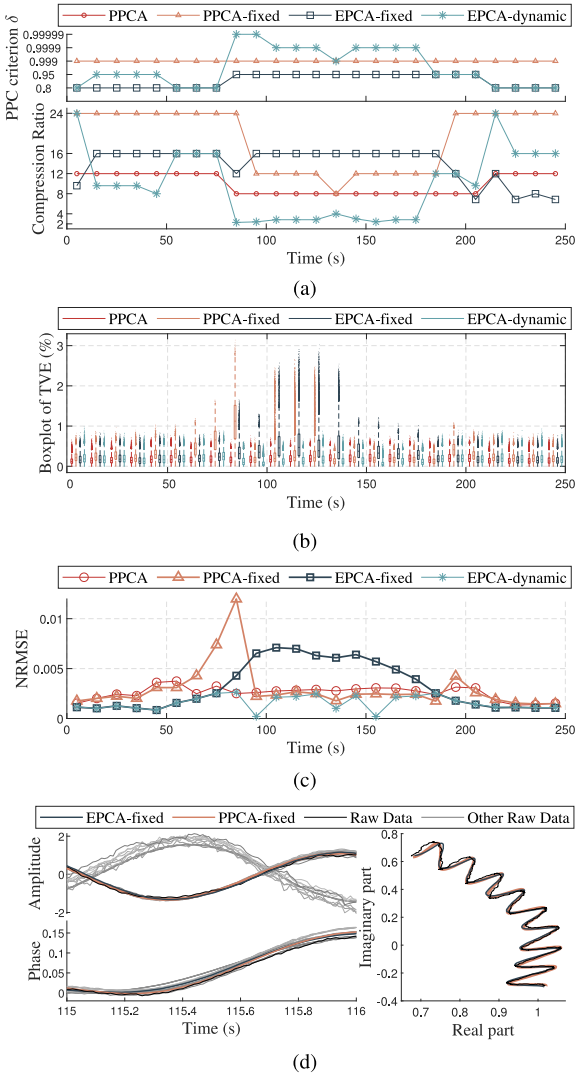


Fig. 7. Compression and reconstruction results of PPCA and EPCA in the LFO incident. (a) Compression ratios in LFO. (b) TVEs of reconstructed data with box plot in LFO. (c) NRMSEs of reconstructed data in LFO. (d) Normalized raw data and "EPCA-fixed" reconstructed data for the A-phase current of G2 that correspond to the TVE outliers in the compression of 110 ∼ 120(s) in (b).

than that of PPCA, the compression ratio of "EPCA-dynamic" is still significantly lower than that of PPCA.

The cause of this different trend for PPCA and EPCA is the selection method of principal components. As stated in Section III, the eigenvalues in $\mathbf{\Lambda}$ correspond to the normalized variances of each principal component in EPCA. When the LFO occurs, the variation of the samples is mainly caused by the disturbance; hence, the similarity features between different separated amplitudes and phases that EPCA extracts is undoubtedly weaker than the coupling relationship between different phasors that PPCA extracts, as explained in Section II-A. Thus, the difference between the normalized variances of each principal component in EPCA, i.e., the eigenvalues, becomes less significant than that in PPCA, as indicated in Table I. Then even $\delta_{\text{criterion}}$ is increased to 0.95 in EPCA, the number of principal components N' does not increase.

TABLE I
EIGENVALUES OF MATRIX **C** IN PCAS FOR THE COMPRESSION OF 110 ∼ 120(S) IN THE LOW FREQUENCY OSCILLATION INCIDENT

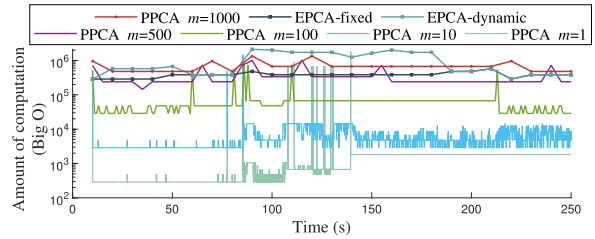
	Phasor	Amplitude	Angle
λ_1	$2.45 \times 10^{+0}$	$4.64 \times 10^{+1}$	$2.40 \times 10^{+1}$
λ_2	1.04×10^{-2}	$3.67 \times 10^{+0}$	6.24×10^{-3}
λ_3	7.29×10^{-4}	9.35×10^{-1}	3.20×10^{-3}
λ_4	4.79×10^{-5}	5.56×10^{-1}	5.07×10^{-4}
N'	$N'_{\text{PPCA}} = 3$ ① $N'_{\text{PPCA-fixed}} = 2$ ②	$N'_{\text{EPCA-fixed}} = 2$ ② $N'_{\text{EPCA-fixed}} = 2$ ③ $N'_{\text{EPCA-dyn.}} = 14$ ②	$N'_{\text{EPCA-fixed}} = 1$ ② $N'_{\text{EPCA-fixed}} = 1$ ③ $N'_{\text{EPCA-dyn.}} = 3$ ②

① Iterative PPC selection, $N' = \min\{N' | \varepsilon_{ij} < \varepsilon_{\text{TVE,MAX}}, \forall i, j\}$
② Normalized cumulative variance criterion, $N' = \min\{N' | \delta \geqslant \delta_{\text{criterion}}\}$
③ Guttman lower bound criterion, $N' = \max\{N' | \lambda_{N'} > 1\}$

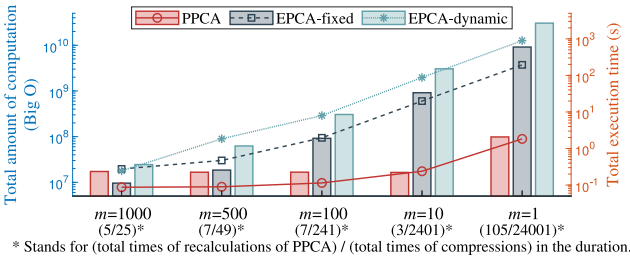
3) Accuracy of the Reconstructed Data: The accuracy of the reconstructed data is illustrated in Fig. 7(b) with the TVEs (12) using the box plot and Fig. 7(c) with the NRMES (19c). The TVEs of the reconstructed data with PPCA and "EPCA-dynamic" are less than 0.8% and 1% throughout the whole 250 s, respectively. Whereas, the TVEs of the reconstructed data with "PPCA-fixed" and "EPCA-fixed" increase significantly during the LFO, significant number of outliers ($\varepsilon_{\text{TVE}} > 2\%$) occur especially. Note that all the medians indicated by the boxplots of the four methods in Fig. 7(b) are still less than 0.5%. During the LFO in 100 ∼ 150(s), the NRMSEs of "PPCA-fixed" are very close to "PPCA", whereas the TVEs of "PPCA-fixed" are not. That is, though the eigenvalue-based criteria are satisfied with limited total errors in "PPCA-fixed", the dynamic snapshots with unacceptably low accuracy occur, as explained in Section III.

To find out the reconstructed dynamic snapshot of phasors with "EPCA-fixed" and "PPCA-fixed" that caused the outliers of boxplot in the compression of 110 ∼ 120(s) in Fig. 7(b), "EPCA-fixed" and "PPCA-fixed" reconstructed data along with the normalized raw data for the A-phase current of G2 are illustrated with colored lines in Fig. 7(d), respectively. Note that these colored data only cause part of the outliers but not all of them. The TVEs of the phasors are significant in Fig. 7(d), as the raw data of the phasor contain lots of distortions and thus causes the outliers of TVE. The reconstructed phasors are much smoother and cannot replicate the distortions of the raw phasors due to the lack of enough PCs. Moreover, the eigenvalues of these methods in the compression of 110 ∼ 120(s) in Figs. 7(b) and 7(c) are listed in Table I. Due to the limitation of the eigenvalue-based criteria, the "PPCA-fixed" and "EPCA-fixed" cannot keep enough PCs to ensure the accuracy; whereas, "EPCA-dynamic" keeps much more PCs than "PPCA" at the sacrifice of compression ratios.

The reason for the inaccurate reconstructed phasors is similar to the reason for the different trends of the compression ratio in Fig. 7(a). Furthermore, the criterion of EPCA and "PPCA-fixed" determines the number of principal components according to eigenvalues λ_i of **C** rather than directly controlling the accuracy of every individual reconstructed data. Whereas, the eigenvalues λ_i depend on the state of power



(a)



(b)

Fig. 8. Amount of computation results of PPCA and EPCA in the LFO incident. (d) Amount of computation for each compression with different methods in LFO. (e) Total amount of computation and execution time under conditions with different m, respectively.

systems even with normalized sample data; thus, it is difficult to determine a uniform threshold $\delta_{\text{criterion}}$ that is suitable for all scenarios of power systems. Consequently, both of the commonly used criteria that use a fixed threshold based on the eigenvalues of covariance matrices may not be suitable for synchrophasor data compression under every circumstance.

4) Amount of Computation and Execution Time: To verify the improvement of the iteration-enhanced PPCA, the illustration of the amount of computation for each compression is shown in Fig. 8(a) with different methods including PPCAs with different m and EPCA, where m is the size of overlapping data windows. "PPCA-fixed" without iterations is not included in this subsection because the amount of computation of it is similar to that of "EPCA-fixed".

The amount of computation of PPCA with $m = 1000$ and EPCAs are at the same order. Meanwhile, as the overlapping data windows reduce from $m = 1000$ to $m = 1$ and the times of compressions increase gradually, the amount of computation of PPCAs is reduced on the order of O(m) accordingly. Moreover, the amount of computation of PPCAs will increase suddenly due to the complete recalculation of PPCA, as indicated in Section IV. The computation required by each recalculation with the iterative PPC selection is determined by the size of overall data window M instead of m; thus, the amount of computation for each recalculation is on the order of O(M), which is the same as EPCAs. These results demonstrate that the complete PPCA with the iterative PPC selection will not cause more intensive computations than EPCAs.

Furthermore, to verify the real-time compression performance, the total amount of computation and execution time of PPCAs during the LFO incident are compared to those of EPCAs that are also enhanced by overlapping data windows, as shown in Fig. 8(b). Compared with "PPCA

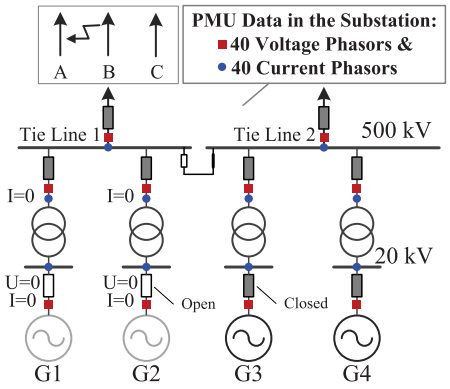


Fig. 9. Actual PMU data of a substation in a two-phase short circuit incident.

$m = 1000$", the computations of "PPCA $m = 1$" have only increased by about 10 times, and these increased computations are mainly due to the increased times of complete PPCA recalculation. That is, the total amount of computation with iteration-enhanced PPCA is on a constant order no matter what m is, as indicated in Section IV. Whereas, compared with the conditions $m = 1000$, the computations of both "EPCA-fixed" and "EPCA-dynamic" with $m = 1$ have increased by more than 1000 times. Also, the increased computations of "EPCA-dynamic" compared to "EPCA-fixed" is caused by more principal components of "EPCA-dynamic".

The trend of the total execution time of each method is similar to that of the total amount of computation, respectively. However, the average execution time of "EPCA-dynamic" with $m = 1$ for each compression is $895.56/24001 = 0.0373$ (s), which is greater than the duration of new data $m/F_s = 0.01$ (s); that is, "EPCA-dynamic" cannot be used for real-time compression with $m = 1$ even enhanced by overlapping data windows.

Consequently, the proposed iteration-based process enhances PPCA with significantly reduced computations and better real-time performance.

B. PPCA in a Two-Phase Short Circuit

In the two-phase short circuit incident, the raw data are from a substation of a hydropower plant with four generators, and the two-phase short circuit fault occurred between A phase and C phase on one of the tie lines as Fig. 9. There are ten measurement points, including two tie lines and four generators with a transformer each. Each measurement point contains the A-phase, B-phase, C-phase, and positive-sequence voltage and current synchrophasors. When the two-phase short circuit incident occurred, G1 and G2 were not in operation and the switches of G1, G2 were open. Thus, the voltages and currents of the buses of G1 and G2 were zero, and the currents of the high-voltage side of the transformers of G1 and G2 were also zero. Consequently, the raw data contain 40 voltage phasors and 40 current phasors. The total time interval for this demonstration is 200 s. The data window of PPCA contains $M = 1000$ samples for each phasor, and the duration T and the sampling rate F_s are 20 seconds and 50 Hz, respectively.

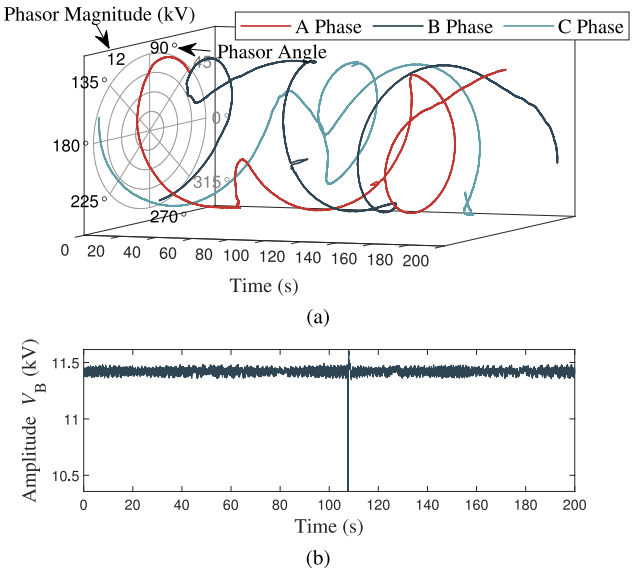


Fig. 10. Raw data of two-phase short circuit incident. (a) Raw data of three-phase voltage synchrophasors in cylindrical coordinates. (b) Raw data of amplitudes of B-phase voltage synchrophasors.

TABLE II
EIGENVALUES OF MATRIX **C** IN PCAS FOR THE COMPRESSION OF $100 \sim 120$(S) IN THE 2-PHASE SHORT CIRCUIT INCIDENT

	Phasor	Amplitude	Angle
λ_1	$1.99 \times 10^{+1}$	$3.25 \times 10^{+2}$	$5.58 \times 10^{+1}$
λ_2	2.01×10^{-1}	$7.16 \times 10^{+1}$	1.74×10^{-1}
λ_3	1.61×10^{-2}	$1.10 \times 10^{+1}$	2.17×10^{-2}
λ_4	6.25×10^{-3}	$9.03 \times 10^{+0}$	5.91×10^{-3}
λ_5	1.88×10^{-3}	$7.44 \times 10^{+0}$	2.89×10^{-3}
λ_6	7.15×10^{-5}	$3.93 \times 10^{+0}$	9.64×10^{-4}
λ_7	3.12×10^{-5}	$2.05 \times 10^{+0}$	2.47×10^{-4}
λ_8	1.42×10^{-5}	$1.31 \times 10^{+0}$	5.71×10^{-5}
λ_9	1.28×10^{-5}	$1.02 \times 10^{+0}$	4.42×10^{-5}
λ_{10}	6.34×10^{-6}	3.12×10^{-1}	3.04×10^{-5}
N'	$N'_{\text{PPCA}} = 9$ ① $N'_{\text{PPCA-fixed}} = 5$ ②	$N'_{\text{EPCA-fixed}} = 4$ ② $N'_{\text{EPCA-fixed}} = 9$ ③ $N'_{\text{EPCA-dyn.}} = 49$ ②	$N'_{\text{EPCA-fixed}} = 1$ ② $N'_{\text{EPCA-fixed}} = 1$ ③ $N'_{\text{EPCA-dyn.}} = 11$ ②

① Iterative PPC selection, $N' = \min\{N' | \varepsilon_{ij} < \varepsilon_{\text{TVE,MAX}}, \forall i, j\}$
② Normalized cumulative variance criterion, $N' = \min\{N' | \delta \geqslant \delta_{\text{criterion}}\}$
③ Guttman lower bound criterion, $N' = \max\{N' | \lambda_{N'} > 1\}$

The demonstrations similar to those in Section V-A are studied and similar results are obtained as shown in Figs. 10 to 12 and Table II. Since the LFO incident in Section V-B is a three-phase symmetrical incident, the features of all phasors are close to each other; however, the features of different phasors will be surely different from each other when an asymmetrical fault occurs. Thus, when the short circuit occurs, the compression ratio of PPCA decreases significantly to 4.42 with $N' = 18$, meaning much more PPCs than those in the LFO. Moreover, the TVEs of "EPCA-fixed" are even over 100% for several synchrophasors during the fault; while, the outliers with $\varepsilon_{\text{TVE}} > 5\%$ of "PPCA-fixed" occur, as shown in Fig. 11(b). Meanwhile, the compression ratio of "EPCA-dynamic" can hardly reach 1.5 to ensure the accuracy of

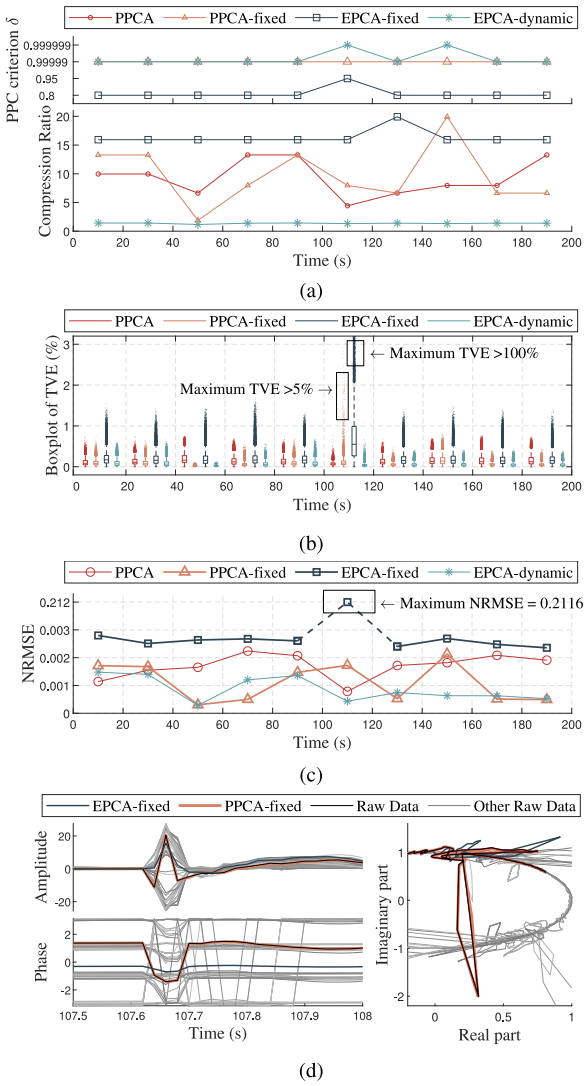


Fig. 11. Compression and reconstruction results of PPCA and EPCA in the two-phase short circuit incident. (a) Compression ratios in two-phase short circuit incident. (b) TVEs of reconstructed data with box plot in two-phase short circuit incident. (c) NRMSEs of reconstructed data in two-phase short circuit incident. (d) Normalized raw data and "EPCA-fixed" reconstructed data for the C-phase current of G3 that correspond to the TVE outliers in the compression of 100 ∼ 120(s) in (b).

reconstructed data, since there are continuous fluctuations in addition to short circuits all the time.

As illustrated in Fig. 11(d), the reconstructed phasors with "EPCA-fixed" are not even close to the raw data with significant errors in both the amplitudes and phases, and "EPCA-fixed" is thus not feasible here. While, though "PPCA-fixed" may track the raw data, it can hardly be used as a result of the unacceptable accuracy of outliers in Fig. 11(b). The cause of this situation is that the trend of C-phase current phasor of G3 is not statistically significant enough to keep the PCs that contain its temporal continuity due to the eigenvalue-based criteria. Moreover, the eigenvalues of these methods in this particular compression of 100 ∼ 120(s) in Figs. 11(b) and 11(d) are also listed in Table II, and the results are similar to those in Table I, even with much more PCs for "EPCA-dynamic". The results of the amount of computation

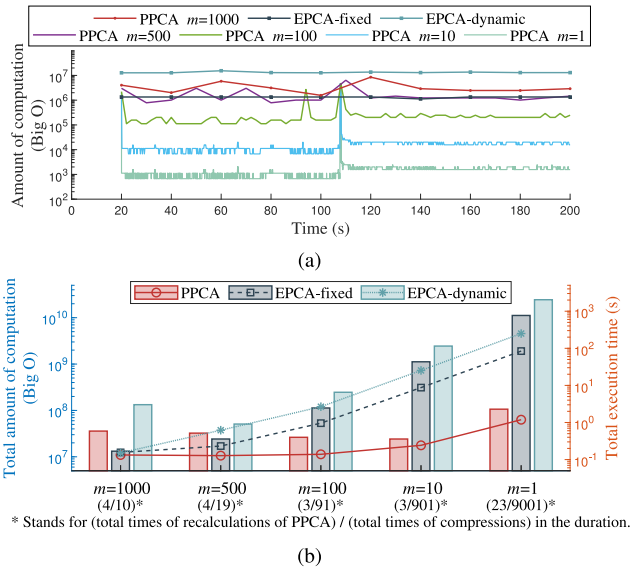


Fig. 12. Results of PPCA and EPCA in the two-phase short circuit incident. (a) Amount of computation for each compression with different methods in two-phase short circuit incident. (b) Total amount of computation and execution time under conditions with different m, respectively.

and execution time in Fig. 12 are similar to those in Fig. 8 in Section V-A.

That is, PPCA still has better performance than EPCA for synchrophasor data compression under asymmetric fault conditions. Whereas, EPCA is not suitable for this situation due to the poor compression ratio or the poor accuracy of reconstructed data, and EPCA even suffers from the computation burden for real-time compressions.

C. Conclusion of Verification

The conclusions of the demonstrations under both the LFO and two-phase short circuit conditions in this section can be drawn as follows.

1) Synchrophasor Data Compression in Complex Numbers: As the disturbance information is provided by the amplitude and phase simultaneously, compressing the amplitude and phase as independent measurements can only use part of the disturbance information; thus, it is more difficult to extract the physically defined spatial and temporal features between different phasors. The proposed PPCA that compresses synchrophasors as a whole in the field of complex numbers can thus achieve better compression performance with both ensured accuracy of reconstructed data and higher compression ratios than compressing amplitudes and phases of a synchrophasor separately.

2) Accuracy of Reconstructed Data: The proposed PPCA directly controls the accuracy of every individual reconstructed data under any conditions by implementing the two-level iterations that take the accuracy of reconstructed data as the criterion for the iterations. Thus, the iteration-enhanced PPCA achieves the balance between the compression ratios and the accuracy of reconstructed data. Whereas, the EPCA and existing PCA-based methods select PCs according to the statistical significance with the eigenvalue threshold criteria;

however, satisfied total errors and unacceptable low accuracy of dynamic snapshots may occur simultaneously under a single circumstance, because the eigenvalue threshold depends on the disturbances seriously and can hardly be selected accurately. This circumstance results in the inevitable and even unacceptable decrease in the accuracy of reconstructed data. The proposed PPCA does not have this problem.

3) Computation Complexity and Execution Time: The iteration-based process with overlapping data windows enhances PPCA by significantly reducing the amount of computation for each compression; thus, the total amount of computation with iteration-enhanced PPCA is on a constant order no matter how frequent the compression is performed. Consequently, PPCA achieves significant real-time performance than conventional PCA-based data compression techniques. Moreover, the iterative PPC selection method does not cause more intensive computations than EPCAs. The iterations in PPCA is feasible and practical since they utilize the physically defined spatial correlation and temporal continuity of phasors. This fact is proved by the results in this section.

VI. Conclusion

To utilize the correlation between amplitudes and phases of synchrophasors for data compression, an iteration-enhanced phasor principal component analysis (PPCA) in the field of complex numbers is proposed to compress synchrophasors as a whole in this article. Two efforts were made to enhance PPCA in addition. First, the proposed PPCA is enhanced by an iterative PPC selection method to ensure the accuracy of reconstructed data without additional computations, since the existing eigenvalue-based criteria are not suitable for data compressions. Second, the proposed PPCA is enhanced by an iteration-based process with overlapping data windows to reduce computations and achieve better real-time performance by taking advantage of the physically defined and rarely-changed spatial correlation between synchrophasors. The proposed iteration-enhanced PPCA is verified with actual PMU data measured under both a low-frequency oscillation incident and a two-phase short circuit incident conditions, respectively. Compared to the EPCA with the eigenvalue-based PC selection, the results demonstrate that PPCA achieves the higher compression ratios with better accuracy of reconstructed data, significantly reduced computations, and better real-time performance under both conditions.

References

[1] F. Zhang, L. Cheng, X. Li, Y. Sun, W. Gao, and W. Zhao, "Application of a real-time data compression and adapted protocol technique for WAMS," *IEEE Trans. Power Syst.*, vol. 30, no. 2, pp. 653–662, Mar. 2015.

[2] M. Cui, J. Wang, J. Tan, A. Florita, and Y. Zhang, "A novel event detection method using PMU data with high precision," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 454–466, Jan. 2019.

[3] M. P. Tcheou *et al.*, "The compression of electric signal waveforms for smart grids: State of the art and future trends," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 291–302, Jan. 2014.

[4] M. Wang *et al.*, "A low-rank matrix approach for the analysis of large amounts of power system synchrophasor data," in *Proc. 48th Hawaii Int. Conf. Syst. Sci.*, Jan. 2015, pp. 2637–2644.

[5] P. H. Gadde, M. Biswal, S. Brahma, and H. Cao, "Efficient compression of PMU data in WAMS," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2406–2413, Sep. 2016.

[6] W. Ren, T. Yardley, and K. Nahrstedt, "ISAAC: Intelligent synchrophasor data real-time compression framework for WAMS," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, 2017, pp. 430–436.

[7] J. E. Tate, "Preprocessing and Golomb–Rice encoding for lossless compression of phasor angle data," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 718–729, Mar. 2016.

[8] R. Klump, P. Agarwal, J. E. Tate, and H. Khurana, "Lossless compression of synchronized phasor measurements," in *Proc. IEEE PES Gen. Meeting*, Jul. 2010, pp. 1–7.

[9] A. Abuadbba, I. Khalil, and X. Yu, "Gaussian approximation-based lossless compression of smart meter readings," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5047–5056, Sep. 2018.

[10] A. Unterweger and D. Engel, "Resumable load data compression in smart grids," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 919–929, Mar. 2015.

[11] T. B. Littler and D. J. Morrow, "Wavelets for the analysis and compression of power system disturbances," *IEEE Trans. Power Del.*, vol. 14, no. 2, pp. 358–364, Apr. 1999.

[12] J. Ning, J. Wang, W. Gao, and C. Liu, "A wavelet-based data compression technique for smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 1, pp. 212–218, Mar. 2011.

[13] N. C. F. Tse, J. Y. C. Chan, W.-H. Lau, J. T. Poon, and L. L. Lai, "Real-time power-quality monitoring with hybrid sinusoidal and lifting wavelet compression algorithm," *IEEE Trans. Power Del.*, vol. 27, no. 4, pp. 1718–1726, Oct. 2012.

[14] J. Khan, S. M. Bhuiyan, G. Murphy, and M. Arline, "Embedded-zerotree-wavelet-based data denoising and compression for smart grid," *IEEE Trans. Ind. Appl.*, vol. 51, no. 5, pp. 4190–4200, Sep./Oct. 2015.

[15] S.-J. Huang and M.-J. Jou, "Application of arithmetic coding for electric power disturbance data compression with wavelet packet enhancement," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1334–1341, Aug. 2004.

[16] J. C. S. de Souza, T. M. L. Assis, and B. C. Pal, "Data compression in smart distribution systems via singular value decomposition," *IEEE Trans. Smart Grid*, vol. 8, no. 1, pp. 275–284, Jan. 2017.

[17] F. Zhang, X. Wang, J. He, Y. Yan, M. Li, and G. Luo, "Phasor data compression with principal components analysis in polar coordinates for subsynchronous oscillations," in *Proc. IEEE Power Energy Soc. General Meeting (PESGM)*, 2019, pp. 1–5.

[18] S. Ryu, H. Choi, H. Lee, and H. Kim, "Convolutional autoencoder based feature extraction and clustering for customer load analysis," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1048–1060, Mar. 2020.

[19] J. Cormane and F. A. de O. Nascimento, "Spectral shape estimation in data compression for smart grid monitoring," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1214–1221, May 2016.

[20] F. A. de O. Nascimento, R. G. Saraiva, and J. Cormane, "Improved transient data compression algorithm based on wavelet spectral quantization models," *IEEE Trans. Power Del.*, vol. 35, no. 5, pp. 2222–2232, Oct. 2020.

[21] G. Lee and Y.-J. Shin, "Multiscale PMU data compression based on wide-area event detection," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, 2017, pp. 437–442.

[22] H. U. Banna, S. K. Solanki, and J. Solanki, "Spatial and temporal redundancy removal in disturbance events recorded by phasor measurement units," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, 2018, pp. 1–5.

[23] R. Mehra, N. Bhatt, F. Kazi, and N. M. Singh, "Analysis of PCA based compression and denoising of smart grid data under normal and fault conditions," in *Proc. IEEE Int. Conf. Electron. Comput. Commun. Technol.*, 2013, pp. 1–6.

[24] J. D. Horel, "Complex principal component analysis: Theory and examples," *J. Clim. Appl. Meteorol.*, vol. 23, no. 12, pp. 1660–1673, Dec. 1984.

[25] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 5th ed. London, U.K.: Sage, 2018.

[26] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[27] *IEEE Standard for Synchrophasor Measurements for Power Systems*, IEEE Standard C37.118.1–2011, Dec. 2011.

[28] A. G. Phadke and T. Bi, "Phasor measurement units, WAMS, and their applications in protection and control of power systems," *J. Modern Power Syst. Clean Energy*, vol. 6, no. 4, pp. 619–629, 2018.

[29] F. Zhang, "Phasor measurement data recorded during low frequency oscillation and short circuit incidents in actual power systems," Dataport, IEEE, Piscataway, NJ, USA, May 2020, doi: 10.21227/1x22-r651.

Fang Zhang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2010 and 2015, respectively. He is currently an Associate Professor with the School of Electrical Engineering, Beijing Jiaotong University, Beijing. His research interests are in the areas of power system dynamic measurement and control based on WAMS, optimal operation, and control of multienergy systems and transportation energy systems.

Xiaojun Wang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from North China Electric Power University, Beijing, China, in 2001, 2004, and 2008, respectively. He is currently a Professor with Beijing Jiaotong University, Beijing. His main research interests are wide-area measurement and multi-energy system operation and control.

Ying Yan (Student Member, IEEE) received the B.S. degree from the Wuhan University of Science and Technology, Wuhan, China, in 2018. She is currently pursuing the graduate degree with the Laboratory of Power System Protection and Control Research, Beijing Jiaotong University. Her research interests are in the areas of power system dynamic measurement.

Jinghan He (Fellow, IEEE) received the B.S. and M.S. degrees from Tianjin University, Tianjin, China, in 1987 and 1994, respectively. She is currently a Professor with Beijing Jiaotong University, Beijing, China. Her main research interests are protective relaying, fault distance measurement, and location in power systems.

Wenzhong Gao (Fellow, IEEE) received the M.S. and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, GA, USA, in 1999 and 2002, respectively. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Denver, Denver, CO, USA. His current research interests include renewable energy, smart microgrid, and WAMS.

Gang Chen (Member, IEEE) was born in 1985. He received the B.S. degree in electrical engineering from Tianjin University, Tianjin, China, in 2008, and the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2013. He is currently a Senior Engineer with the State Grid Sichuan Electric Power Research Institute, Chengdu, China. His research interests include power system dynamic monitoring and control and low- and ultra-low-frequency oscillation analysis and control.