Data-Driven Power Flow Linearization: A Regression Approach

Article *in* IEEE Transactions on Smart Grid · February 2018

DOI: 10.1109/TSG.2018.2805169

CITATIONS
99

READS
1,202

5 authors, including:

Yuxiao Liu
Tsinghua University
17 PUBLICATIONS **179** CITATIONS
SEE PROFILE

Ning Zhang
Tsinghua University
254 PUBLICATIONS **6,282** CITATIONS
SEE PROFILE

Yi Wang
The University of Hong Kong
155 PUBLICATIONS **4,429** CITATIONS
SEE PROFILE

Jingwei Yang
Tsinghua University
22 PUBLICATIONS **1,314** CITATIONS
SEE PROFILE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid

Data-Driven Power Flow Linearization: A Regression Approach

Yuxiao Liu, *Student Member, IEEE,* Ning Zhang, *Member, IEEE,* Yi Wang, *Student Member, IEEE,* Jingwei Yang, *Student Member, IEEE,* and Chongqing Kang, *Fellow, IEEE*

***Abstract*—The linearization of a power flow (PF) model is an important approach for simplifying and accelerating the calculation of a power system's control, operation, and optimization. Traditional model-based methods derive linearized PF models by making approximations in the analytical PF model according to the physical characteristics of the power system. Today, more measurements of the power system are available and thus facilitate data-driven approaches beyond model-driven approaches. This work studies a linearized PF model through a data-driven approach. Both a forward regression model ((P,Q) as a function of (θ,V)) and an inverse regression model ((θ,V) as a function of (P,Q)) are proposed. Partial least squares (PLS)- and Bayesian linear regression (BLR)-based algorithms are designed to address data collinearity and avoid overfitting. The proposed approach is tested on a series of IEEE standard cases, which include both meshed transmission grids and radial distribution grids, with both Monte Carlo simulated data and public testing data. The results show that the proposed approach can realize a higher calculation accuracy than model-based approaches can. The results also demonstrate that the obtained regression parameter matrices of data-driven models reflect power system physics by demonstrating similar patterns with some power system matrices (e.g., the admittance matrix).**

***Index Terms*—Power flow, linearization, data-driven, least squares regression, Bayesian inference**

I. Introduction

Power flow (PF) analysis is the basis of power system analysis and optimization. The nonlinearity of PF equations leads to difficulties in optimization and control algorithms [1], as in non-convergence problems, and incurs high computation burdens. The linearization of PF equations can markedly simplify the complexity and ensure the convergence of algorithm, which is why it is already widely used in power system control [2], [3], scheduling [4] and market clearing [5]-[7]. Among all of the PF linearization approaches, the DC power flow (DCPF) equations are currently the most widely used in industry. DCPF reveals the approximated linearity relationship between active power injection (P) and phase angle (θ). A substantial number of studies has been conducted to enhance the DCPF and have considered the formulation of reactive power injections (Q) and voltage magnitudes (V) [8]-[12]. In [8], PF equations are formulated as a linearized form with respect to the square of the voltage magnitude (V^2) and modified phase angle ($V^2\theta$). On this basis, linearized models have been further proposed using the square of the voltage magnitude and phase angle [9], the decoupled voltage magnitude and phase angle [10, 11], and the logarithmic transform of voltage magnitude and phase angle [12] as the independent variable. The above linearization methods improve the accuracy beyond DCPF.

With the spread of massive phasor measurement units (PMUs) and supervisory control and data acquisition (SCADA) systems, measurement data from power systems are sufficient to be used in rebuilding system models. The methods are known as data-driven methods and are found to contribute to the efficiency and accuracy of power system analysis. Traditionally in power system PF analysis, many model-based approaches based on a precise PF model are used to derive the models that facilitate rapid calculation or ensure optimization convergence. We consider these approaches to be model-to-model approaches. Different from model-to-model approaches, many data-driven methods rediscover model parameters from various operation data, which is denoted as the data-to-model approach. Chen *et al* proposed a measurement-based method to estimate the distribution factors [13] and the Jacobian matrix [14] using the least squares method. The rediscovered model parameters have advantages in near real-time flexibility to adapt to changes in topology or load. Yuan *et al.* identified the admittance matrix in a distribution network using graph theory [15]. The model can recover the real-time topology and admittance matrix in distribution grids with several hidden buses without measurements. Few works have focused on the non-network-parameter-based data-driven method of PF calculations. To the best knowledge of the authors, the closest work is [16], which uses non-linear support vector regression (SVR) to reveal relationships among variables in a PF analysis. A nonlinear mapping rule between active and reactive bus injection (P,Q) and the phase angle and voltage magnitude (V,θ) is built based on historical data. However, the phase angle and voltage magnitude are considered in a coupling form ($\phi(V,\theta)$), which cannot consider different bus types in the mapping process. For example, the *PV* bus cannot be considered for the coupling of the phase angle and voltage magnitude.

Our work focuses on the data-driven linearization method for PF analysis. Compared with the current model-based PF

This work was supported in part by the National Key R&D Program of China (No. 2016YFB0900100), the National Science Foundation of China (No. 51620105007), and the Scientific & Technical Project of State Grid: theoretical and empirical research of the key technology for the whole process management of power grid operation risk based on multi source data mining. (Corresponding author: Chongqing Kang.)

Y. Liu, N. Zhang, Y. Wang, J. Yang, and C. Kang are with the State Key Lab of Power Systems, Department of Electrical Engineering, Tsinghua University, Beijing 100084, China. (E-mail: cqkang@tsinghua.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid

linearization approaches mentioned above, the data-driven linearization PF analysis has the following advantages: 1) It does not require knowledge of the system topologies and parameters. In distribution grids, due to frequent re-configurations and increasing penetration of distributed energy resources, the exact system topologies, element parameters, and the control logic of active control devices are difficult to model accurately [16]-[18]. Data-driven approaches are merely based on historical measurements and thus have significant advantages under these circumstances. 2) It improves the linearization accuracy of PF calculations. The training data reflects the real operation status of the power system such that the parameters of the data-driven approach more accurately consider the power system operation condition than model-based approaches do. For example, the data-driven approach can consider the deviation of parameters due to the atmospheric condition and aging [19].

Compared with the current data-driven approaches for PF analysis mentioned in [13]-[15], the proposed method has the following advantages: 1) Reducing calculation errors. The current data-driven approach identifies the parameters in the PF model first and then uses the identified model to conduct further control and optimizations. The PF calculation error may accumulate in the data-to-model and model-to-data processes. Our work replaces the process with a direct data-to-data approach and thus avoids modeling errors. 2) Reducing the computational burden. The current data-driven model still obtains a non-linearized PF model [16] and suffers from the computational burden for further applications, such as probabilistic load flow [20]. The proposed method reveals the linearized mapping relationship between operation variables based on historical data. 3) Enhancing computational flexibility. The proposed data-driven method can flexibly solve PF problems by considering different settings on bus types.

The contributions of this paper can be summarized as follows:

1) A data-driven linearization approach of PF equations is proposed that does not require knowledge of the power grid parameters and considers PF physics.

2) Both forward and inverse regression models of PF equations are produced that facilitates PF calculations with different settings of bus types.

3) Both partial least squares (PLS)- and Bayesian linear regression (BLR)-based algorithms are introduced to address the collinearity and avoid the overfitting of real operation data.

The remainder of this paper is organized as follows: Section II revisits the PF linearization problem from a data perspective and provides the framework of the data-driven PF linearization approach. Section III proposes forward and inverse models to regress linearized the parameters of PF equations. Section IV introduces PLS and BLR models. Section V validates the accuracy and robustness of the proposed model on several cases. In Section VI, we present further discussions, including the net loss considerations, the required training data size, and the potential applications and challenges of the proposed method. Finally, conclusions are drawn in Section VII.

II. Data-Driven Power Flow Linearization Framework

This section first interprets the PF calculation from a data-driven point of view. Then, the idea of the data-driven PF linearization is presented with discussions on its feasibility and challenges. The framework of data-driven PF linearization is finally proposed.

A. PF Calculation from a Data-Driven Perspective

The steady state of a power system can be uniquely described by the power injections, bus voltage, and branch power flow. The measurements of these quantities at a certain snapshot (e.g., at time t) can be formulated as a vector $\boldsymbol{x}_t$. The expression of $\mathbf{x}_t$ is shown in (1), where P_i^t, Q_i^t, V_i^t and θ_i^t represent the active power injection, reactive power injection, voltage magnitude and voltage angle of the bus i at time t, respectively; PF_i^t and QF_i^t represent the active and reactive power flow of branch l at time t, respectively.

$$\begin{aligned}\mathbf{x}_t = (&P_1^t, \ \dots, \ P_N^t, \ Q_1^t, \ \dots, \ Q_N^t, \\ &V_1^t, \ \dots, \ V_N^t, \ \theta_1^t, \ \dots, \ \theta_N^t, \\ &PF_1^t, \ \dots, \ PF_N^t, \ QF_1^t, \ \dots, \ QF_N^t, \ \dots)\end{aligned} \tag{1}$$

Vector $\boldsymbol{x}_t$ represents a power system operation snapshot in time t. From the data point of view, it can be seen as a high-dimensional vector in a high-dimensional hyperspace. All of the vectors $\mathbf{x}_t$ that describe different operation states of a power system are on the high-dimensional surface described by the nonlinear AC power flow (ACPF) equations in (2):

$$\begin{aligned}P_i &= V_i \sum_{j\in i} V_j (G_{ij}\cos\theta_{ij} + B_{ij}\sin\theta_{ij}) \\ Q_i &= V_i \sum_{j\in i} V_j (G_{ij}\sin\theta_{ij} - B_{ij}\cos\theta_{ij}) \\ PF_{ij} &= (V_i^2 - V_iV_j\cos\theta_{ij})g_{ij} - V_iV_j\sin\theta_{ij}b_{ij} \\ QF_{ij} &= -(V_i^2 - V_iV_j\cos\theta_{ij})b_{ij} - V_iV_j\sin\theta_{ij}g_{ij}\end{aligned} \tag{2}$$

where G_{ij} and B_{ij} represent the real and imaginary parts of the i th row and j th column of the admittance matrix respectively, and PF_{ij} and QF_{ij} represent the active and reactive branch flows from bus i to bus j respectively.

From a data-driven point of view, PF equations can be built using a data mining technique instead of an admittance matrix. In this work, the identification of the parameters in the PF equations can be seen as a multi-parameter regression for a high-dimensional surface based on historical operation data. The obtained PF equations from the regression can be further used in the PF calculation, control or operation in the same way as traditional model-based PF equations.

Current studies on the linearization of PF equations show that even though the expression of PF equations is non-linear, it has a high degree of linearity such that the linearized model does not lose too much accuracy. Therefore, we can deduce that the high-dimensional surface described by the ACPF can be approximated as a hyperplane. Though it will introduce

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid

errors, the linearity of the model would provide better performance on the computational speed and convergence in further power flow applications.

B. PF Linearization Visualization

To visualize the linearization of PF equations in a hyperplane, a simple two-bus system is illustrated, as shown in Fig. 1, where the PF equations can be shown in three-dimensional space.

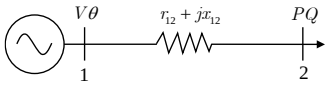


Fig. 1. An illustrative two-bus system

We compare the ACPF with two representative linearized PF models: 1) traditional DCPF equations and 2) the decoupled linear power flow (DLPF) equations proposed in [10]. The formulation of DLPF is shown in (3), where $\mathbf{B}'$ represents the imaginary part of the admittance matrix without shunt elements.

$$\begin{bmatrix}\mathbf{P}\\ \mathbf{Q}\end{bmatrix}=-\begin{bmatrix}\mathbf{B}' & -\mathbf{G}\\ \mathbf{G} & \mathbf{B}\end{bmatrix}\begin{bmatrix}\boldsymbol{\theta}\\ \mathbf{V}\end{bmatrix} \tag{3}$$

In this two-bus system, there are only two PF equations and two independent variables. The independent variables include the active and reactive power injection of bus #2 P_2 and Q_2. The dependent variables include the voltage magnitude and angle of bus #2 V_2 and θ_2. Bus #1 is the reference bus. Fig. 2 shows the value of θ_2 with different combinations of P_2 and Q_2, calculated by ACPF, DLPF, and DCPF, respectively. Three conclusions can be observed from Fig. 2:

1) The non-linear ACPF surface has a high degree of linearity in the two-bus system, which suggests that linear regression is worth to be examined in PF calculation.

2) The approximation of DLPF is closer to ACPF than the approximation of DCPF is.

3) The two model-based linear approximations (DCPF and DLPF) still result in clear errors, which suggests that the data-driven linearization approaches still have much to improve on with regards to accuracy.

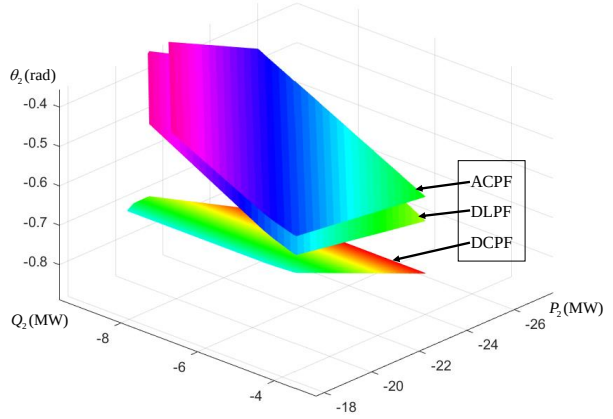


Fig. 2. Visualization of ACPF, DCPF, and DLPF in a two-bus system

C. PF Mapping Directions

In this paper, the linearization between P,Q and θ,V is considered. Other relationships (e.g., between P,Q, and V^2 [9], $V^2\theta$ [8] and $\ln V$ [12]) are also available and can be similarly handled.

To explore the linearized mapping rule between P,Q, and θ,V, our first attempt is to regress the parameters mapping from θ,V to P,Q, mathematically:

$$P,Q=f(\theta,V) \tag{4}$$

This direction is consistent with the mapping direction of the ACPF equations in (2), where the function of P,Q with respect to θ,V has an explicit expression. We name this type of mapping direction forward regression.

We also consider the mapping direction from P,Q to θ,V. Such a mapping direction is in accordance with the procedure of the PF calculation, where P,Q are known and θ,V are to be calculated. We name this mapping direction inverse regression.

D. Challenges of Regression

The challenges of such a regression lay in two main aspects: to address the collinearity of data and to avoid overfitting. First, collinearity among the voltage angle and magnitude data is inevitable because of the similar rise and fall patterns among the different buses [21]. This will result in ill-conditioned regression and larger errors of PF calculation. Second, the number of variables in the regression parameter matrices for large power systems may be far greater than the amount of historical operation data that represents the current system situation. Such a characteristic may incur the problem of overfitting. Although the performance of an overfitted model may perform surprisingly well on the training dataset, the accuracy may suffer a great loss on the testing dataset [22]. To overcome these two challenges, a PLS-based regression is proposed to ensure the calculation accuracy under collinearity, and a BLR-based regression is proposed to avoid the overfitting of the regressed parameters.

E. Framework of Data-driven PF Linearization

Based on the above discussions, a framework for data-driven PF linearization is proposed in Fig. 3. The framework is divided into three parts: linearization models, regression methods, and data types.

First, two linearization models, the forward regression model and the inverse regression model, are proposed for different purposes for the PF calculation. Second, both PLS- and BLR-based regression methods are established for both regressions. Finally, different data types, random data under a Monte Carlo simulation and data with collinearity (loads in different buses rise and fall in similar patterns) from the public dataset, are tested to illustrate the validity of the proposed models. Each part is detailed in the following sections.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid

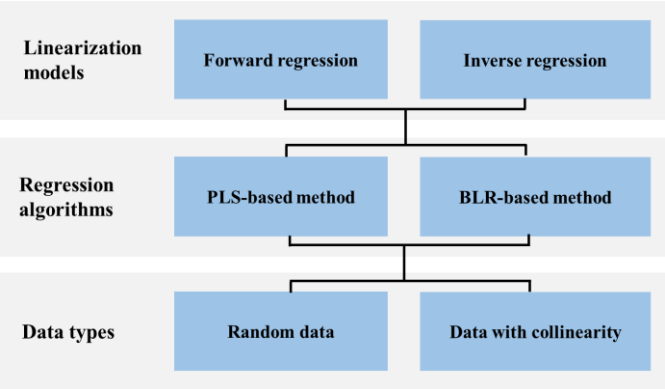


Fig. 3. Framework of data-driven power flow linearization

III. Power Flow Linearization Models

In this section, the models of forward regression and inverse regression are formulated. A theoretical derivation is conducted to reveal the relationship between the regressed matrices and several power system matrices.

A. Forward Regression Model

The generalized linearization equations of the forward regression model are shown in (5), where $\mathbf{C}_P$ and $\mathbf{C}_Q$ are constant terms of active and reactive bus injections.

$$\begin{bmatrix}\mathbf{P}\\ \mathbf{Q}\end{bmatrix}=\begin{bmatrix}\mathbf{H} & \mathbf{N}\\ \mathbf{M} & \mathbf{L}\end{bmatrix}\begin{bmatrix}\boldsymbol{\theta}\\ \mathbf{V}\end{bmatrix}+\begin{bmatrix}\mathbf{C}_P\\ \mathbf{C}_Q\end{bmatrix} \tag{5}$$

Although the ACPF equations do not have any constant terms, $\mathbf{C}_P$ and $\mathbf{C}_Q$ are added to the linearization equations to enhance the regression capability of the model. In power system operation, the value of some independent variables may remain unchanged, and the regression parameters of these independent variables in H, N, M, and L may not be regressed. The influences of these independent variables can be absorbed in this constant terms.

The potential application of forward regression is introduced in [16]. It can be used in the PF analysis of a distribution grid more accurately than the traditional model-based PF calculation can because the former considers the invisible control actions of active controllers in the distribution network by learning from historical data.

B. Inverse Regression Model

It is proposed that inverse regression can calculate PF when considering different bus types, e.g., $PQ, PV, V\theta$ buses. The known and unknown variables in the PF calculation are different among different types of buses. Moreover, the bus types may change from one to another during the calculation process. Our goal is to find the regression model that can obtain the mapping of all of the known variables to the unknown variables for various conditions.

We arrange different types of buses in the following sequence: $PQ, PV, V\theta$.

$$\begin{aligned}&\mathbf{P}=\begin{bmatrix}\mathbf{P}_L^T & \mathbf{P}_S^T & \mathbf{P}_R^T\end{bmatrix}^T \quad \mathbf{Q}=\begin{bmatrix}\mathbf{Q}_L^T & \mathbf{Q}_S^T & \mathbf{Q}_R^T\end{bmatrix}^T\\ &\mathbf{V}=\begin{bmatrix}\mathbf{V}_L^T & \mathbf{V}_S^T & \mathbf{V}_R^T\end{bmatrix}^T \quad \boldsymbol{\theta}=\begin{bmatrix}\boldsymbol{\theta}_L^T & \boldsymbol{\theta}_S^T & \boldsymbol{\theta}_R^T\end{bmatrix}^T\end{aligned} \tag{6}$$

The inverse regression equations can be expressed as a block matrix form in (7), where $\mathbf{C}_1 \sim \mathbf{C}_6$ are constant terms and $\mathbf{A}_{ij}$ is the regression parameter matrix. It should be noted that during the regression stage, both $\begin{bmatrix}\boldsymbol{\theta}_L & \boldsymbol{\theta}_S & \mathbf{P}_R & \mathbf{V}_L & \mathbf{V}_S & \mathbf{V}_R\end{bmatrix}^T$ and $\begin{bmatrix}\mathbf{P}_L & \mathbf{P}_S & \mathbf{Q}_L & \mathbf{Q}_S & \mathbf{Q}_R\end{bmatrix}^T$ are known, and $\mathbf{A}_{ij}$ and $\mathbf{C}_1 \sim \mathbf{C}_6$ are parameters that need to be estimated.

$$\begin{bmatrix}\boldsymbol{\theta}_L\\ \boldsymbol{\theta}_S\\ \mathbf{P}_R\\ \mathbf{V}_L\\ \mathbf{V}_S\\ \mathbf{V}_R\end{bmatrix}=\left[\begin{array}{ccccc}\mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} & \mathbf{A}_{14} & \mathbf{A}_{15}\\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & \mathbf{A}_{24} & \mathbf{A}_{25}\\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} & \mathbf{A}_{34} & \mathbf{A}_{35}\\ \mathbf{A}_{41} & \mathbf{A}_{42} & \mathbf{A}_{43} & \mathbf{A}_{44} & \mathbf{A}_{45}\\ \hdashline \mathbf{A}_{51} & \mathbf{A}_{52} & \mathbf{A}_{53} & \mathbf{A}_{54} & \mathbf{A}_{55}\\ \mathbf{A}_{61} & \mathbf{A}_{62} & \mathbf{A}_{63} & \mathbf{A}_{64} & \mathbf{A}_{65}\end{array}\right]\begin{bmatrix}\mathbf{P}_L\\ \mathbf{P}_S\\ \mathbf{Q}_L\\ \mathbf{Q}_S\\ \mathbf{Q}_R\end{bmatrix}+\begin{bmatrix}\mathbf{C}_1\\ \mathbf{C}_2\\ \mathbf{C}_3\\ \mathbf{C}_4\\ \mathbf{C}_5\\ \mathbf{C}_6\end{bmatrix} \tag{7}$$

When obtaining all of the parameters in $\mathbf{A}_{ij}$ and $\mathbf{C}_1 \sim \mathbf{C}_6$, the PF calculations can be conducted. When calculating the PF, $\boldsymbol{\theta}_L$, $\boldsymbol{\theta}_S$, $\mathbf{P}_R$ and $\mathbf{V}_L$ in (7) are unknown variables, whereas $\mathbf{V}_S$ and $\mathbf{V}_R$ are known variables. Similarly, as for the independent variables, $\mathbf{P}_L$, $\mathbf{P}_S$ and $\mathbf{Q}_L$ are known, whereas $\mathbf{Q}_S$ and $\mathbf{Q}_R$ are unknown. Hence, the equation in (7) can be rewritten in the form of (8), where $\mathbf{x}_1 = [\mathbf{P}_L, \mathbf{P}_S, \mathbf{Q}_L]^T$ and $\mathbf{y}_2 = [\mathbf{V}_S, \mathbf{V}_R]^T$ are known variables, and $\mathbf{x}_2 = [\mathbf{Q}_S, \mathbf{Q}_R]^T$ and $\mathbf{y}_1 = [\boldsymbol{\theta}_L, \boldsymbol{\theta}_S, \mathbf{P}_R, \mathbf{V}_L]^T$ are unknown variables. After obtaining all of the parameters from the regression, the unknown variables can be calculated in (9). The invertibility of matrix $\tilde{\mathbf{A}}_{22}$ is discussed in later sections.

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix}=\begin{bmatrix}\tilde{\mathbb{A}}_{11} & \tilde{\mathbb{A}}_{12}\\ \tilde{\mathbb{A}}_{21} & \tilde{\mathbb{A}}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}+\begin{bmatrix}\tilde{\mathbb{C}}_1\\ \tilde{\mathbb{C}}_2\end{bmatrix} \tag{8}$$

$$\begin{aligned}\mathbf{x}_2 &= \tilde{\mathbb{A}}_{22}^{-1}(\mathbf{y}_2 - \tilde{\mathbb{A}}_{21}\mathbf{x}_1 - \tilde{\mathbb{C}}_2)\\ \mathbf{y}_1 &= \tilde{\mathbb{A}}_{11}\mathbf{x}_1 + \tilde{\mathbb{A}}_{12}\mathbf{x}_2 + \tilde{\mathbb{C}}_1\end{aligned} \tag{9}$$

The reasons for building the regression equation in (7) in such form are twofold:

1) To maintain the feasibility of the PF calculation with flexible bus type settings.

The model of inverse regression is applicable for different bus type settings because all buses are reordered and calculated in the same sequence of $PQ, PV, V\theta$. When the bus types transform from one into another, the regression parameter matrix $\mathbf{A}_{ij}$ in (7) is reordered rather than recalculated.

The necessary condition of such ability comes from the fact that $\tilde{\mathbb{A}}_{22}$ in (8) is reversible. In the inverse regression model, the independent variables corresponding to $\tilde{\mathbb{A}}_{22}$ are reactive power injections of the PV and $V\theta$ buses. The dependent variables corresponding to $\tilde{\mathbb{A}}_{22}$ are the voltage magnitude of the PV and $V\theta$ buses. These quantities are not constants in the historical data. Regarding the voltage magnitude, the fluctuation of PV and $V\theta$ buses is inevitable. The maximum fluctuation range of each bus depends on the voltage control device (e.g., the maximum range is set as 0.05p.u.-0.215p.u. in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid

continental Europe [23]). Therefore, the obtained $\tilde{\mathbb{A}}_{22}$ is a full ranked matrix. In contrast, matrices $\mathbf{H},\mathbf{N},\mathbf{M},\mathbf{L}$ in (5) obtained from forward regression cannot be used to obtain the mapping via the formulation of (7)-(9). This is because the $\tilde{\mathbb{A}}_{22}$ in the forward regression corresponds to the dependent variables of the *PQ* and *PV* buses. When there are zero-power-injected *PQ* buses, the regression cannot obtain a full ranked or nonsingular $\tilde{\mathbb{A}}_{22}$, so the mapping can hardly be obtained.

2) To remove $\mathbf{P}_R$ from independent variables to avoid collinearity.

In (7), all the active and reactive power injections are independent variables, except the active power injection of the $V\theta$ bus $\mathbf{P}_R$, because the relationship of active injection of all buses can be approximated in (10).

$$\sum_i P_i \approx 0 \tag{10}$$

In other words, $\mathbf{P}_R$ can be almost determined by active power injections of other buses. The regression will lead to collinearity and will result in the ill-conditioned regression parameter matrix $\mathbf{A}_{ij}$. Although the relation in (10) is not a strict equation, it results in the problem of collinearity. For more on the impact of collinearity on the regression, refer to [24], [25]. Instead, the regression model for $\mathbf{P}_R$ is added as the third row in (7). Such a formulation can consider the power balance of the power system.

C. Relationship with Physical Parameter Matrices

Interestingly, the value of forward and inverse regression parameter matrices is numerically similar to the value of several power system matrices. The derivation process of this relationship is shown in Fig. 4.

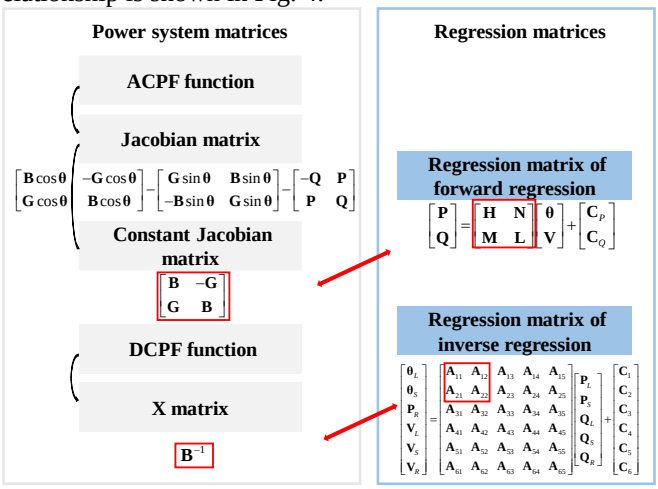


Fig. 4. Derivation process of relationships between regression parameter matrices and several power system matrices

Fig. 4 presents a theoretical derivation of how the regression parameter matrices are related to several power system matrices. First, the forward regression can be seen as the first-order Taylor's approximation of the PF equations in (2). Therefore, the forward regression parameter matrix can be seen as the partial derivative of the PF equations: the Jacobian matrix. However, given a set of power system operation data, the value of the Jacobian matrix is different for different operating points, whereas the value of the regression parameter matrix is constant. Hence, the constant Jacobian matrix that is widely used in the Newton-Raphson method is a reasonable approximation of the forward regression parameter matrix.

Second, it is complicated to derive the theoretical explanation of the inverse regression parameter matrix from the partial derivative of the PF equations because $\boldsymbol{\theta}$, $\mathbf{V}$ are difficult to represent as a function of $\mathbf{P}$, $\mathbf{Q}$ with definite formulations. Thus, the derivation of the inverse partial derivative (e.g., $\partial\boldsymbol{\theta}/\partial\mathbf{P}$, $\partial\boldsymbol{\theta}/\partial\mathbf{Q}$, $\partial\mathbf{V}/\partial\mathbf{P}$, $\partial\mathbf{V}/\partial\mathbf{Q}$) from the PF equations require implicit differentiation. From the four inverse partial derivatives, $\partial\boldsymbol{\theta}/\partial\mathbf{P}$ can be easily approximated by the inverse matrix of $\mathbf{B}$ according to the DCPF equations:

$$\mathbf{P} = \mathbf{B}\boldsymbol{\theta} \tag{11}$$

The approximation of $\partial\boldsymbol{\theta}/\partial\mathbf{P}$ corresponds to the matrices $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$ and $\mathbf{A}_{22}$ in (7). These relationships discussed above can serve as an indicator of overfitting.

D. Mapping of Branch Power Flow

The mapping of the branch PF is similar to the mapping of the power injection. Given the historical data of active and reactive branch PF (*PF* and *QF*), the mapping rule can be regressed. The mapping direction can either be from P,Q to PF,QF or from θ,V to PF,QF. Taking the former as an example, the linearization equations is in (12), where $\mathbf{P}_R$ is removed from the independent variables:

$$\begin{bmatrix}\mathbf{PF}\\ \mathbf{QF}\end{bmatrix} = \begin{bmatrix}\mathbf{H}_{line} & \mathbf{N}_{line}\\ \mathbf{M}_{line} & \mathbf{L}_{line}\end{bmatrix}\begin{bmatrix}\mathbf{P}\\ \mathbf{Q}\end{bmatrix} + \begin{bmatrix}\mathbf{C}_{PF}\\ \mathbf{C}_{QF}\end{bmatrix} \tag{12}$$

IV. Regression Algorithms

The mathematical models of the forward regression, inverse regression, and branch PF regression are linear regression models. To simplify the representation, the generalized regression equation is presented by $\mathbf{A}$, $\mathbf{X}$ and $\mathbf{Y}$, which represent the regression parameter matrix, matrix of independent variables and matrix of dependent variables, respectively.

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \tag{13}$$

In the data-driven PF regression, $\mathbf{X}$ and $\mathbf{Y}$ are matrices rather than vectors. The number of columns represents different datasets at different times (or different operation snapshots in a power system), and the number of rows represents different variables. Taking forward regression as an example, $\mathbf{X}$ and $\mathbf{Y}$ are formulated in (14), where the last row of $\mathbf{X}$ corresponds to the constant terms.

$$\mathbf{X} = \begin{bmatrix}\boldsymbol{\theta}^1 \dots \boldsymbol{\theta}^t \dots \boldsymbol{\theta}^T\\ \mathbf{V}^1 \dots \mathbf{V}^t \dots \mathbf{V}^T\\ 1 \dots 1 \dots 1\end{bmatrix}, \mathbf{Y} = \begin{bmatrix}\mathbf{P}^1 \dots \mathbf{P}^t \dots \mathbf{P}^T\\ \mathbf{Q}^1 \dots \mathbf{Q}^t \dots \mathbf{Q}^T\end{bmatrix} \tag{14}$$

In the process of theoretical derivation, $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{Y}$ are expressed in the form of rows:

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_N\end{bmatrix}^T \mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2 & \dots & \mathbf{y}_M\end{bmatrix}^T \mathbf{A} = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_M\end{bmatrix}^T \tag{15}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid

In the following sub-section, both PLS-based and BLR-based algorithms are proposed.

A. PLS-Based Algorithm

The objective of the PLS-based algorithm is to regress between two zero-mean data blocks, $N \times T$ matrix $\mathbf{X}$ and $M \times T$ matrix $\mathbf{Y}$. It can address the collinearity and lack of observations after combining the features from principal component analysis (PCA) and canonical correlation analysis (CCA) [26]. PLS decomposes p components from the matrix of independent variables $\mathbf{X}$ and matrix of dependent variables $\mathbf{Y}$ into the form

$$\begin{aligned} \mathbf{X} &= \mathbf{C}\mathbf{T}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{R}\mathbf{U}^T + \mathbf{F} \end{aligned} \tag{16}$$

where $\mathbf{T}, \mathbf{U}$ are $T \times p$ matrices of the p extracted components, $\mathbf{C}, \mathbf{R}$ are $N \times p$ and $M \times p$ matrices represent the loading matrices, and $\mathbf{E}, \mathbf{F}$ are $N \times T$ and $M \times T$ matrices and represent the residuals. The datasets of $\mathbf{X}$ and $\mathbf{Y}$ share a similar rise and fall pattern in different rows, which corresponds to the collinearity in the power system data (e.g., active power injections tend to rise and fall at the same time in different buses). In (16), PLS projects $\mathbf{X}$ and $\mathbf{Y}$ onto two small matrices $\mathbf{T}$ and $\mathbf{U}$ to extract the key components that $\mathbf{Y}$ correlate to $\mathbf{X}$.

Calculation of the correlated matrices is based on the nonlinear iterative partial least squares algorithm [24]. Finally, given the matrices of $\mathbf{X}^*$ and $\mathbf{Y}^*$ as the updated independent and dependent variables, the matrix of dependent variables is predicted in the form

$$\mathbf{Y}^* = \mathbf{A}\mathbf{X}^* \ \textit{where}\ \mathbf{A}^T = \mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} \tag{17}$$

B. BLR-based Algorithm

The BLR-based algorithm is conducted within the context of Bayesian inference [27]. Different vectors of $\mathbf{Y}$ in (15) are regressed. Taking $\mathbf{y}_i$ as an example, (18) represents the regression equation

$$\mathbf{y}_i = \mathbf{a}_i\mathbf{X} + \mathbf{e}_i,\ i = 1, 2, ..., M \tag{18}$$

where $\mathbf{e}_i$ represents the additive noise of $\mathbf{y}_i$, and $\mathbf{X}$, $\mathbf{y}_i$ are centered in a previous step. Each $\mathbf{a}_i$ represents a vector:

$$\mathbf{a}_i = \begin{bmatrix} a_{i1} \ ... \ a_{ij} \ ... \ a_{iL} \end{bmatrix} \tag{19}$$

According to the Bayesian inference framework, the posterior probability of $\mathbf{a}_i$ follows

$$p(\boldsymbol{a}_i \mid \boldsymbol{y}_i, \boldsymbol{X}) \propto p(\boldsymbol{a}_i)\, p(\boldsymbol{y}_i, \boldsymbol{X} \mid \boldsymbol{a}_i) \tag{20}$$

where $p(\boldsymbol{a}_i)$ represents the prior and $p(\boldsymbol{y}_i, \boldsymbol{X} \mid \boldsymbol{a}_i)$ represents the likelihood. The prior is introduced to avoid overfitting by setting a simple form of presumption on the posterior distribution. In this work, an elliptical Gaussian distribution prior for $\mathbf{a}_i$ is assumed:

$$p(\boldsymbol{a}_i) = \prod_{j=0}^{L} N(a_{ij} \mid 0, \beta_j^{-1}) \tag{21}$$

The knowledge of power flow model is embedded into prior of the BLR model from two aspects. 1) The prior distribution is set to be zero centered. In Section III.C, the regression parameter matrix of forward regression is demonstrated to be sparse. The zero centered prior can be regarded as a form of regularization. It will promote sparsity in forward regression and reduce parameter overfitting in both forward and inverse regression. 2) The prior distribution is set to be elliptical because the precision may vary among different parameters. In other words, each a_j has its own standard deviation β_j^{-1} not predefined so that it can adjust according to the real observation data. The reciprocal of standard deviations $\boldsymbol{\beta}$ follow a Gamma distribution, which can be determined by maximizing the marginal likelihood $p(\boldsymbol{y}_i \mid \boldsymbol{X}, \boldsymbol{\beta})$ [28]. Under the assumption of a Gaussian distribution with additive noise $\mathbf{e}_i$, the likelihood can be written as

$$p(\boldsymbol{y}_i, \boldsymbol{X} \mid \boldsymbol{a}_i) = (2\pi\beta_j^{-2})^{-L/2} \exp\{-\frac{\beta_j^2}{2}\left\|\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{a}_i\right\|^2\} \tag{22}$$

To calculate parameter $\boldsymbol{a}_i$, a maximum a posterior (MAP) optimization is conducted. During the iteration of the optimization process, the estimation of a_j is set to zero when its deviation β_j^{-1} is under a certain threshold. This gives a flexibility of adjustment on the sparsity. The reasonable sparsity is essential for forward regression whereas the inverse regression does not share this characteristic. Therefore, the threshold is set as a large number in forward regression and a small number in inverse regression. The detailed derivation, optimization process and proving of convergence are in reference [28].

V. Experimental Results

A. Data Generation

Data of the power system operation measurement used in the case studies are divided into two categories: Monte Carlo simulation and public testing data. Data processing from both categories is performed in MATLAB with the aid of MATPOWER 6.0 [29]. Parameters are regressed using the training dataset, and the PF calculation accuracy is tested using the newly generated testing dataset. The source code is available online [30].

1) Monte Carlo simulation

A Monte Carlo simulation was run on meshed transmission grids, which include IEEE 5, 30, 57, and 118-bus systems and radial distribution grids, which include the IEEE 33-bus system [31] and the modified 123-bus system [32]. The active load consumption is calculated from the preset load consumption multiplied by a factor randomly drawn from a uniform distribution over the interval [0.8, 1.2]. The reactive load consumption is calculated from the active load consumption multiplied by a factor randomly drawn from a uniform distribution over the interval [0.15, 0.25].

2) Public testing data

We use the hourly load data of the NREL-118 test system [33] to replace the randomly generated active load consumption in the Monte Carlo simulation. The load data are synthetic that have similar rise and fall patterns that were

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid

TABLE I. ERRORS OF FORWARD, INVERSE AND BRANCH CALCULATION ON DIFFERENT CASES

Cases	Size of training data	Size of testing data	Forward calculation					Inverse calculation				Branch PF calculation				
			Errors	DCPF	DLPF	PLS	BLR	Errors	DLPF	PLS	BLR	Errors	DCPF	DLPF	PLS	BLR
IEEE 5	100	300	P	24.11	1.117	**0.412**	0.615	θ	0.020	**8.2e-4**	6.8e-3	PF	8.120	8.120	**0.126**	0.609
			Q	---	66.21	**0.940**	1.065	V	7.8e-4	**2.0e-5**	4.1e-3	QF	---	---	**8.934**	256.1
IEEE 30	100	300	P	12.49	0.578	**0.034**	0.238	θ	0.154	**1.9e-3**	0.071	PF	7.734	7.562	**0.104**	0.825
			Q	---	12.66	**0.404**	0.471	V	9.9e-4	**1.0e-5**	1.4e-3	QF	---	---	**1.340**	226.9
IEEE 33	100	300	P	67.05	1.114	**0.012**	**0.012**	θ	0.028	**4.3e-4**	0.011	PF	1.142	1.142	**5.0e-3**	8.8e-3
			Q	---	0.759	0.044	**0.027**	V	2.0e-3	**7.3e-6**	6.5e-4	QF	---	---	**0.013**	0.497
IEEE 57	300	300	P	98.11	7.343	**0.262**	2.132	θ	0.215	**0.036**	0.218	PF	19.16	13.22	**0.395**	0.965
			Q	---	26.83	**0.300**	2.990	V	7.1e-3	**2.1e-4**	1.1e-3	QF	---	---	**5.227**	24.71
IEEE 118	300	300	P	16.89	4.546	**0.061**	1.385	θ	2.593	**0.074**	0.296	PF	86.96	86.04	**2.263**	7.078
			Q	---	77.85	**1.096**	31.73	V	1.9e-3	**1.2e-4**	8.1e-4	QF	---	---	**5.570**	68.27
NREL 118	300	300	P	85.90	9.486	**0.161**	1.207	θ	3.003	0.622	**0.271**	PF	33.08	29.37	10.59	**4.326**
			Q	---	107.4	**0.486**	3.982	V	2.3e-3	**6.3e-4**	7.6e-4	QF	---	---	**28.07**	36.53
Modified 123	300	300	P	12.49	0.512	**0.007**	0.452	θ	0.091	**3.2e-4**	2.6e-3	PF	0.319	0.319	**5.1e-4**	6.6e-3
			Q	---	2.071	**0.003**	**0.003**	V	2.3e-3	**3.2e-6**	3.5e-6	QF	---	---	**3.6e-3**	7.4e-3

✧ The errors of P, Q, PF, and QF are in mean absolute percentage error with the unit of 100%, whereas the errors of θ and V are in mean absolute error.
✧ The errors of Q correspond to DCPF are not shown because DCPF cannot calculate reactive power. The errors of QF that correspond to DCPF and DLPF are not shown because DCPF and DLPF cannot calculate the reactive power flow.

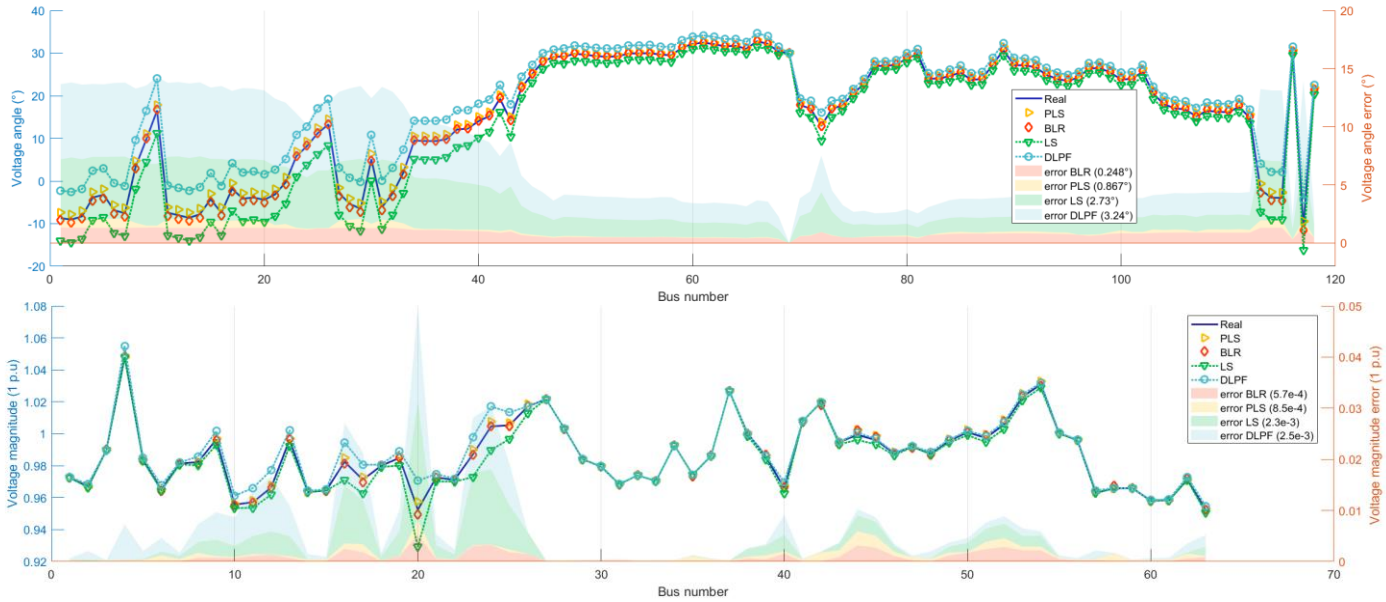


Fig. 5. Calculation results of voltage angles and magnitudes of each bus for NREL-118 test system using different method

learned from 1980-2012 weather and load data. Gaussian noise is added to the original load data because the original load data is divided on a pro-rata basis at different times. It is a plausible method for insufficient datasets [34]. A scale factor is multiplied by the load data to balance the system generation capacity. The data of the NREL-118 test system are used to test the model adaptability to the data with collinearity.

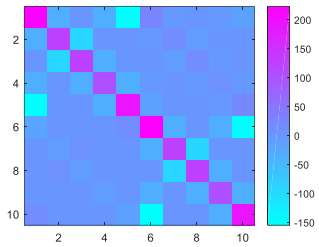
B. Basic Results

We first fit the data-driven PF equations on the training dataset using the proposed regression algorithms. Then, the PF calculation is conducted on the testing dataset. The average errors of the proposed PLS- and BLR-based algorithms compared with both the DCPF and DLPF methods on forward calculation, inverse calculation, and branch PF calculation are shown in Table I, where the size of training or testing data represents the number of snapshots. Each snapshot is represented in the form of (1).

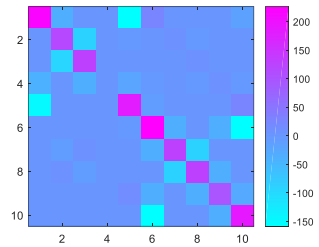
The accuracy of forward regression is measured by the errors of power injection P and Q in all buses; the accuracy of inverse regression is measured by the errors of voltage magnitude V in PQ buses and voltage angle θ in all buses except the $V\theta$ bus. As is illustrated in Table I, several conclusions can be drawn:

1) Among all cases of PF calculations, the proposed data-driven approaches are more accurate than or at least as accurate as that of model-based DCPF and DLPF methods.

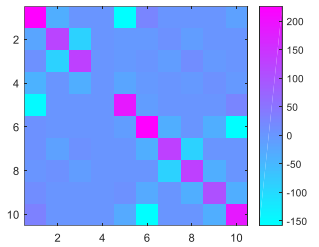
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid



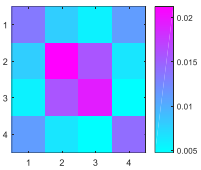
Constant Jacobian matrix



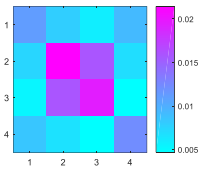
Forward regression parameter matrix under the PLS-based algorithm



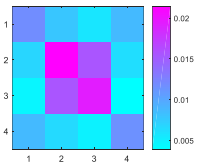
Forward regression parameter matrix under the BLR-based algorithm



The inverse matrix of **B** in DCPF

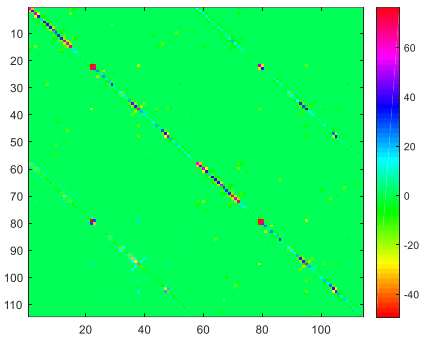


Inverse regression parameter matrix under the PLS-based algorithm

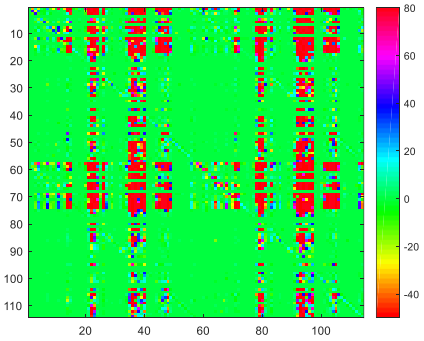


Inverse regression parameter matrix under the BLR-based algorithm

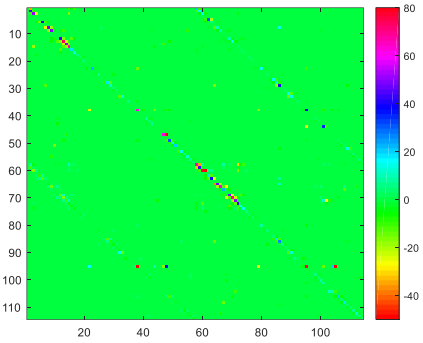
Fig. 7(a). Comparisons between regression parameter matrices and several power system matrices of IEEE 5-bus system



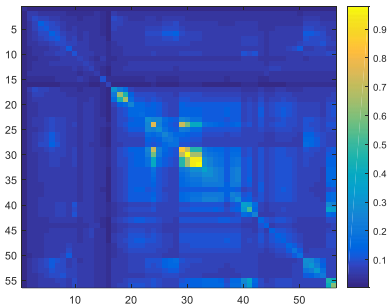
Constant Jacobian matrix



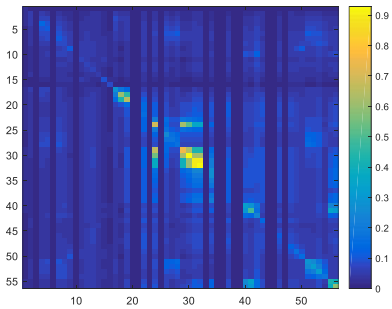
Forward regression parameter matrix under the PLS-based algorithm



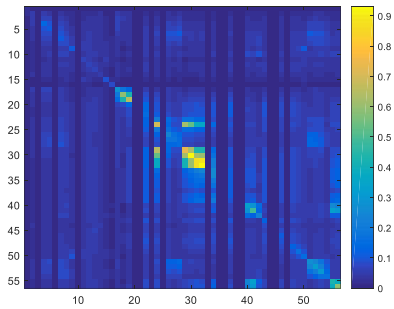
Forward regression parameter matrix under the BLR-based algorithm



The inverse matrix of **B** in DCPF



Inverse regression parameter matrix under the PLS-based algorithm



Inverse regression parameter matrix under the BLR-based algorithm

Fig. 7(b). Comparisons between regression parameter matrices and several power system matrices of IEEE 57-bus system

2) The proposed data-driven approaches provide better results in areas where model-based methods are not as accurate, such as reactive power injection Q of forward calculations and the active power flow PF. The errors in these areas are one order less than that of model-based methods.

3) In most of the cases (IEEE 5, 30, 57, 118, modified 123-bus systems), the PLS-based algorithm is more accurate than the BLR-based algorithm. The BLR-based algorithm only demonstrates better results on IEEE 33-bus in P,Q calculations and NREL-118 test systems in θ, PF calculations.

C. Calculation Results under Data Collinearity

The inverse calculation results of the NREL-118 test system using different methods are illustrated in Fig. 5 and Fig. 6 to present both detailed and overall computation accuracy. Fig. 5 shows the calculation errors of voltage angle and magnitude for a certain snapshot among all the 300 testing snapshots. The accuracy of voltage magnitude only accounts for the PQ buses. To test the algorithm effectiveness under data collinearity, the least squares (LS) regression that does not consider data collinearity is conducted for comparison. It is clear that the proposed algorithm performs better than both the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid

LS and the DLPF algorithms, in terms of both the voltage angle and the voltage magnitude calculations, in almost every bus of the NREL-118 test system. The results of all 300 snapshots are illustrated by the histograms in Fig. 6, where the error of every snapshot is the average error of all buses of the NREL-118 test system. We can conclude from Fig. 6 that the proposed algorithms have smaller errors than that of the LS and the DLPF algorithms. The errors of LS algorithm are distributed in a wide range in both voltage angle and voltage magnitude calculations, showing that its performance is not stable under data collinearity. The results demonstrate that the regression algorithms do require some special considerations on data collinearity. Fig. 5 and Fig. 6 validate the effectiveness of the proposed regression algorithms under data collinearity.

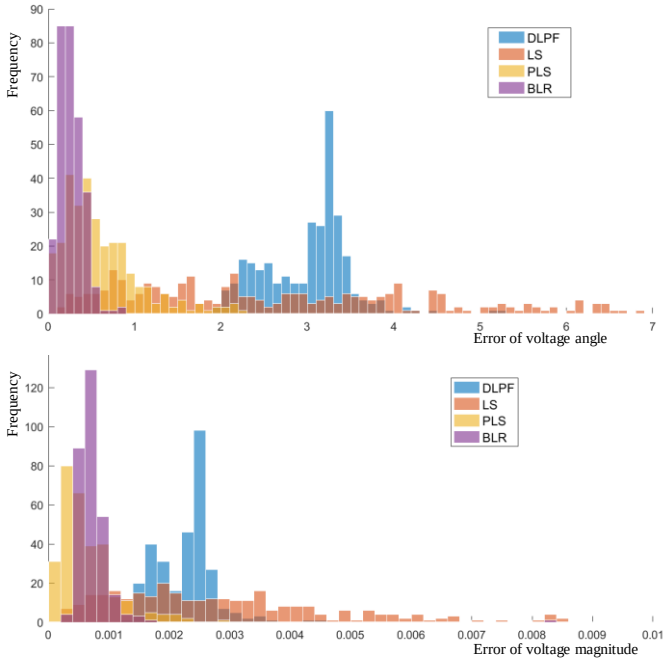


Fig. 6. Histograms of calculation errors of voltage angles and magnitudes for NREL-118 test system

D. Regression Parameters

To verify the relationship between regression parameter matrices and several power system matrices, IEEE 5 and 57-bus systems are analyzed. The forward regression parameter matrix of the IEEE 5-bus system based on the PLS and BLR-based algorithms compared with the constant Jacobian matrix are shown in Fig. 7(a). Regarding the inverse regression parameter matrix, $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$ and $\mathbf{A}_{22}$ in (7) compared with the inverse matrix of $\mathbf{B}$ in DCPF are also shown in Fig. 7(a). Similarly, the same comparisons of the IEEE 57-bus system are shown in Fig. 7(b). Regarding the IEEE 5-bus system, the forward regression parameter matrices of both the PLS- and BLR-based algorithms are extremely similar to the constant Jacobian matrix. The $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$ and $\mathbf{A}_{22}$ in the inverse regression of both the PLS and BLR-based algorithms are similar to the inverse matrix of $\mathbf{B}$ in DCPF. These results validate the theoretical analysis in Fig. 4. Regarding the IEEE 57-bus system, the constant Jacobian matrix is highly sparse with diagonal non-zero parameters. The forward regression parameter matrix of the PLS-based algorithm contains the diagonal non-zero parameters and many other off-diagonal large value parameters. This parameter overfitting can be eliminated in the BLS-based algorithm. The inverse matrix of $\mathbf{B}$ in DCPF is a full matrix; thus, the regression parameter matrices of both PLS and BLR-based algorithms provide an acceptable approximation. There are several zero columns because of the zero-power-injected buses. The columns that correspond to these buses are regressed to zero and have no influence on the calculation accuracy as long as the injections of these buses remain zero.

VI. Discussion

A. Net Loss

The model proposed in this paper considers net loss in the form of first order approximation.

The active and reactive net losses of branch connecting bus i and bus j are represented in (23):

$$\begin{aligned} LP_{ij} &= g_{ij}(V_i^2 - 2V_iV_j\cos\theta_{ij} + V_j^2) \\ LQ_{ij} &= -b_{ij}(V_i^2 - 2V_iV_j\cos\theta_{ij} + V_j^2) \end{aligned} \tag{23}$$

where LP_{ij} and LQ_{ij} represent the active and reactive line losses, respectively. The net losses can be calculated by the summation of branch flow from bus i to j and bus j to i:

$$\begin{aligned} LP_{ij} &= PF_{ij} + PF_{ji} \\ LQ_{ij} &= QF_{ij} + QF_{ji} \end{aligned} \tag{24}$$

After solving (5) for forward regression or (7) for inverse regression, the net losses are given by (12) and (24). The calculation of P, Q in (5) and (7) is the linear combination of voltage angles, voltage magnitudes, and constant terms, where the net losses are embedded in the first order approximation form. Thereafter, the calculation of PF, QF in (12) is the linear combination of active and reactive power injections and constant terms, where the parameters are optimized to minimize the errors. This approximation is proved to have an acceptable accuracy of active net losses in Reference [35]. The quantitative study on the net losses of the data-driven linearization model could be further explored in future work.

B. Size of Training Data

The issue of required training data size is of great importance in a regression-based method. Fig.8 demonstrates the change of the voltage angle errors with respect to the size of training data in different cases. The results are calculated based on the inverse regression using PLS-based algorithm.

Generally, more training data is required for larger system since larger system has more independent variables in power flow model and more parameters needs to be regressed. Hence, the horizontal axis of Fig. 8 is set as the ratio of training data to the number of buses. It can be observed in Fig. 8 that the results become stable when the ratio is more than 2.4. Note that the number of independent variables is twice as much as the number of buses (see (5) or (7)). That means the minimum required training data size is only around 1.2 times as much as the number of independent variables.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2018.2805169, IEEE Transactions on Smart Grid

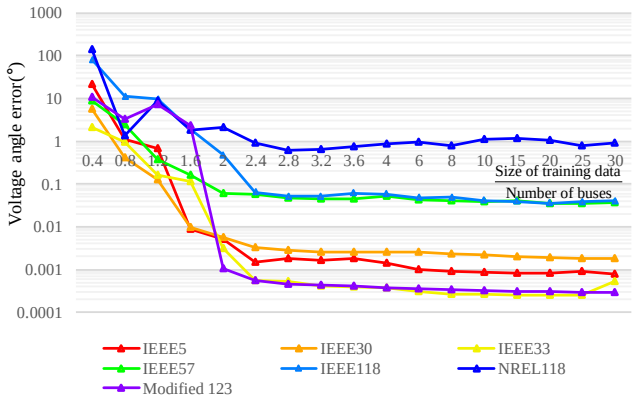


Fig. 8. Change of the voltage angle errors with respect to the size of training data in different cases

This interesting result can be explained by the negative correlation between the squared multiple correlation coefficient and the minimum training data size [36]. Furthermore, Milton deduced a training data size formula based on the F-test for significance of a regression coefficient [37]:

$$N = K + 1 + \frac{t_{stat}^{\ 2}(1-R^2)}{\Delta R_k^{\ 2}} \qquad (25)$$

where N is the training data size, K is the number of independent variables, t_{stat} is the t-statistic in F-test, R^2 is the squared multiple correlation coefficient, and $\Delta R_k^{\ 2}$ is the explained variance attributed to the k^{th} variable when entered last in the regression equation in the F-test. The statistic results of these variables in different cases are listed in Table II. We set $t_{stat} = 3$ under the assumption that the p-value in F-test is less than 0.01 [34], which is a conditional term flexibly set by the researchers.

TABLE II. VARIABLES IN THE TRAINING DATA SIZE FORMULA IN DIFFERENT CASES

Cases	R^2	$\Delta R_k^{\ 2}$	$\frac{t_{stat}^{\ 2}(1-R^2)}{\Delta R_k^{\ 2}}$	$\frac{N}{K}$
IEEE5	0.999468	3.653e-02	0.35	1.14
IEEE30	0.999962	3.137e-03	0.12	1.02
IEEE33	0.999790	9.386e-04	3.37	1.07
IEEE57	0.999783	9.763e-04	2.05	1.03
IEEE118	0.999991	2.569e-04	0.35	1.01
NREL118	0.999997	1.728e-06	17.91	1.08
IEEE123	0.999608	1.955e-02	0.18	1.00

In Table II, the ratio of training data to the number of buses (N/K) in all cases are less than 1.2 times the number of independent variables, which theoretically validates the empirical analysis in Fig. 8.

The reason that the N/K ratio is only slightly more than one is that the squared multiple correlation coefficient (R^2) is extremely close to one. That means an extremely high proportion of the variance in the dependent variables can be predicted from the independent variables. The reasons are twofold: 1) The ACPF equations can be well described by linear equations. This high degree of linearity is visualized in a two-bus system in Fig. 2. 2) The variables in power systems have definite numerical relationships described by the ACPF equations. This is different from the regression analysis applied in many other statistic researches without definite numerical relationships between dependent variables and independent variables. 3) The linearized equations only need to approximate the power flow near operation state of the power system rather than all the mathematical feasible state described by the ACPF equations.

Still, the required training data size of PF regression is a new research field and should be further investigated by more theories and numerical studies.

C. Potential Applications and Challenges

The proposed linearization method provides an accurate and efficient approach in power system analysis that requires a large amount of PF calculations. For example, the probabilistic load flow needs repeated power flow calculation of numerous sampled power system states to consider the uncertainty of renewable energy and loads. It has the issue of the balance between calculation speed and accuracy. Our method produces a fast and accurate linearized PF calculation approach that can be utilized in probabilistic load flow problems. Moreover, the proposed method provides linear power flow formulations to enhance the convergence in PF related optimization problems. The ACPF equations introduce nonlinear constraints in power system optimization problems (e.g. the optimal power flow), making it an ordinary nonlinear programming optimization problem with convergence issue. The linearized PF equations in our approach can serve as an alternative to the nonlinear ACPF constraints and thus achieve a desirable convergence performance. The method can be used in power system operation, e.g. in security constraint economic dispatch, local marginal price calculation, and etc., the same way as traditional model-based PF equations can be used.

Furthermore, the proposed method enables the PF calculation in situations where the exact system topologies, element parameters, and the control logic are difficult to obtain. Such situation widely exists in distribution grids with active control devices or with high penetrations of renewable energy.

However, it is currently challenging for the proposed method when the system topologies should be frequently changed. For instance, in the $N-1$ contingency analysis, the outages of transmission lines lead to the change of power system matrices. It cannot be regressed with data under normal operating conditions. Future research will focus on exploring the data-driven methods considering frequent topology changes.

VII. CONCLUSION

In this paper, we provide a data-driven PF linearization approach that bridges the gap between model-based PF linearization methods and data-driven power system analysis approaches. Forward and inverse regression methods as well as branch PF mapping are proposed to facilitate a variety of linearized PF calculation. To conquer the collinearity of the data, PLS- and BLR-based regression methods are used. Several cases, including meshed transmission grids, radial distribution grids, and public testing system, are examined. The results verify the distinct advantage on the accuracy of the

proposed data-driven approaches over several selected methods. More importantly, the parameter matrices obtained from the regression are found to maintain physical significance of the model-based parameters, which demonstrates its ability to identify the physical reality of the power system.

We envision that the proposed data-driven linearization approach serves as the foundation of accurate linearization calculations and optimization methods.

REFERENCES

[1] B. Stott, J. Jardim and O. Alsac, "DC Power Flow Revisited," *IEEE Trans. Power Syst.*, vol. 24, pp. 1290-1300, 2009.

[2] P. Kundur, N. J. Balu and M. G. Lauby, Power system stability and control vol. 7: McGraw-hill New York, 1994.

[3] L. An, X. U. Qianming, M. A. Fujun, and C. Yandong, "Overview of power quality analysis and control technology for the smart grid," *Journal of Modern Power Systems and Clean Energy*, vol. 4, pp. 1-9, 2016.

[4] J. Liu, M. Kazemi, A. Motamedi, H. Zareipour, and J. Rippon, "Security-Constrained Optimal Scheduling of Transmission Outages with Load Curtailment," *IEEE Trans. Power Syst.*, p. 1-1, 2017.

[5] H. Yuan, F. Li, Y. Wei, and J. Zhu, "Novel Linearized Power Flow and Linearized OPF Models for Active Distribution Networks with Application in Distribution LMP," *IEEE Trans. Smart Grid*, p. 1-1, 2016.

[6] M. Chávez-Lugo, C. R. Fuerte-Esquivel, C. A. Cañizares, and V. J. Gutierrez-Martinez, "Practical Security Boundary-Constrained DC Optimal Power Flow for Electricity Markets," *IEEE Trans. Power Syst.*, vol. 31, pp. 3358-3368, 2016.

[7] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Science and Technology*, vol. 20, pp. 117-129, 2015.

[8] S. M. Fatemi, S. Abedi, G. B. Gharehpetian, S. H. Hosseinian, and M. Abedi, "Introducing a Novel DC Power Flow Method With Reactive Power Considerations," *IEEE Trans. Power Syst.*, vol. 30, pp. 3012-3023, 2015.

[9] Z. Yang, H. Zhong, A. Bose, T. Zheng, Q. Xia, and C. Kang, "A Linearized OPF Model with Reactive Power and Voltage Magnitude: A Pathway to Improve the MW-Only DC OPF," *IEEE Trans. Power Syst.*, p. 1-1, 2017.

[10] J. Yang, N. Zhang, C. Kang, and Q. Xia, "A State-Independent Linear Power Flow Model with Accurate Estimation of Voltage Magnitude," *IEEE Trans. Power Syst.*, vol. 22, pp. 3607-3617, 2017.

[11] Y. Wang, N. Zhang, H. Li, J. Yang, and C. Kang, "Linear three-phase power flow for unbalanced active distribution networks with PV nodes," *CSEE Journal of Power and Energy Systems*, vol. 3, pp. 321-324, 2017.

[12] Z. Li, J. Yu, Q. H. Wu, " Approximate Linear Power Flow Using Logarithmic Transform of Voltage Magnitudes with Reactive Power and Transmission Loss Consideration," *IEEE Trans. Power Syst.*, p. 1-1, 2017.

[13] Y. C. Chen, A. D. Dominguez-Garcia and P. W. Sauer, "Measurement-based estimation of linear sensitivity distribution factors and applications," *IEEE Trans. Power Syst.*, vol. 29, pp. 1372-1382, 2014.

[14] Y. C. Chen, J. Wang, A. D. Domínguez-García, and P. W. Sauer, "Measurement-based estimation of the power flow Jacobian matrix," *IEEE Trans. Power Syst.*, vol. 7, pp. 2507-2515, 2016.

[15] Y. Yuan, O. Ardakanian, S. Low, and C. Tomlin, "On the inverse power flow problem," *arXiv preprint arXiv:1610.06631*, 2016.

[16] J. Yu, Y. Weng and R. Rajagopal, "Mapping Rule Estimation for Power Flow Analysis in Distribution Grids," *arXiv preprint arXiv:1702.07948*, 2017.

[17] Y. Weng, Y. Liao and R. Rajagopal, "Distributed Energy Resources Topology Identification via Graphical Modeling," *IEEE Trans. Power Syst.*, vol. 32, pp. 2682-2694, 2017.

[18] D. Zhang, J. Li and D. Hui, "Coordinated control for voltage regulation of distribution network voltage regulation by distributed energy storage systems," *Protection and Control of Modern Power Systems*, vol. 3, p. 3, 2018.

[19] S. Gajare, A. K. Pradhan and V. Terzija, "A Method for Accurate Parameter Estimation of Series Compensated Transmission Lines Using Synchronized Data," *IEEE Trans. Power Syst.*, p. 1-1, 2017.

[20] P. Zhang and S. T. Lee, "Probabilistic load flow computation using the method of combined cumulants and Gram-Charlier expansion," *IEEE Trans. Power Syst.*, vol. 19, pp. 676-682, 2004.

[21] J. Zhang, L. Guan and C. Y. Chung, "Instantaneous sensitivity identification in power systems-challenges and technique roadmap," in *Power and Energy Society General Meeting (PESGM)*, Boston, MA, 2016, pp. 1-5.

[22] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, pp. 1-12, 2004.

[23] O. A. Mousavi and R. Cherkaoui, "Literature survey on fundamental issues of voltage and reactive power control," *Ecole Polytechnique Fédérale de Lausanne: Lausanne, Switzerland*, 2011.

[24] D. A. Belsley, E. Kuh and R. E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity* vol. 571: John Wiley & Sons, 2005.

[25] E. Cheney and D. Kincaid, *Numerical mathematics and computing*: Nelson Education, 2012.

[26] R. Rosipal and N. Mer, "Overview and recent advances in partial least squares," in *International Conference on Subspace, Latent Structure and Feature Selection*, 2005, pp. 34-51.

[27] G. E. Box and G. C. Tiao, *Bayesian inference in statistical analysis* vol. 40: John Wiley & Sons, 2011.

[28] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, pp. 211-244, 2001.

[29] R. D. Zimmerman, C. E. Murillo-Sanchez and R. J. Thomas, "MATPOWER: Steady-State Operations, Planning, and Analysis Tools for Power Systems Research and Education," *IEEE Trans. Power Syst.*, vol. 26, pp. 12-19, 2011.

[30] Source code for the data-driven power flow linearization, Github. [Online]. Available: https://github.com/YuxiaoLiu/data-driven-power-flow-linearization

[31] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Trans. Power Del.*, vol. 4, no. 2, pp. 1401-1407, Apr. 1989.

[32] S. Bolognani and S. Zampieri, "On the Existence and Linear Approximation of the Power Flow Solution in Power Distribution Networks," *IEEE Trans. Power Syst.*, vol. 31, pp. 163-172, 2014.

[33] I. Pena, C. Brancucci and B. M. Hodge, "An Extended IEEE 118-bus Test System with High Renewable Penetration," *IEEE Trans. Power Syst.*, p. 1-1, 2017.

[34] J. Xie, T. Hong, T. Laing, and C. Kang, "On normality assumption in residual simulation for probabilistic load forecasting," IEEE Trans. Smart Grid, vol. 8, pp. 1046-1053, 2017.

[35] B. Eldridge, R. O'Neill, A. Castillo, "An Improved Method for the DCOPF with Losses," *IEEE Trans. Power Syst.*, p. 1-1, 2017.

[36] G. T. Knofczynski and D. Mundfrom, "Sample sizes when using multiple linear regression for prediction," *Educational and Psychological Measurement*, vol. 68, pp. 431-442, 2008.

[37] S. Milton, "A sample size formula for multiple regression studies," *Public Opinion Quarterly*, vol. 50, pp. 112-118, 1986.

Yuxiao Liu (S'16) received the B.S. degree from the Electrical Engineering Department of Tsinghua University in China in 2016.

He is currently pursuing Ph.D. degree in Tsinghua University. His research interests include data-driven methods in smart grid.

Ning Zhang (S'10-M'12) received both a B.S. and Ph.D. from the Electrical Engineering Department of Tsinghua University in China in 2007 and 2012, respectively.

He is currently an Associate Professor at the same university. His research interests include multiple energy system integration, renewable energy, and power system planning and operation.

Yi Wang (S'14) received the B.S. degree from the Department of Electrical Engineering in Huazhong University of Science and Technology (HUST), Wuhan, China, in 2014.

He is currently pursuing Ph.D. degree in Tsinghua University. He is also a visiting student researcher at the University of Washington, Seattle, WA, USA. His research interests include data analytics in smart grid and multiple energy systems.

Jingwei Yang (S'15) received the B.S. degree from the Electrical Engineering Department of Tsinghua University in China in 2015.

He is currently pursuing Ph.D. degree in Tsinghua University. His research interests include optimal power flow, renewable energy and multiple energy system integration.

Chongqing Kang (M'01-SM'08-F'17) received the Ph.D. degree from the Department of Electrical Engineering in Tsinghua University, Beijing, China, in 1997.

He is currently a Professor in Tsinghua University. His research interests include power system planning, power system operation, renewable energy, low carbon electricity technology and load forecasting.