Big Data Technology and its Applications

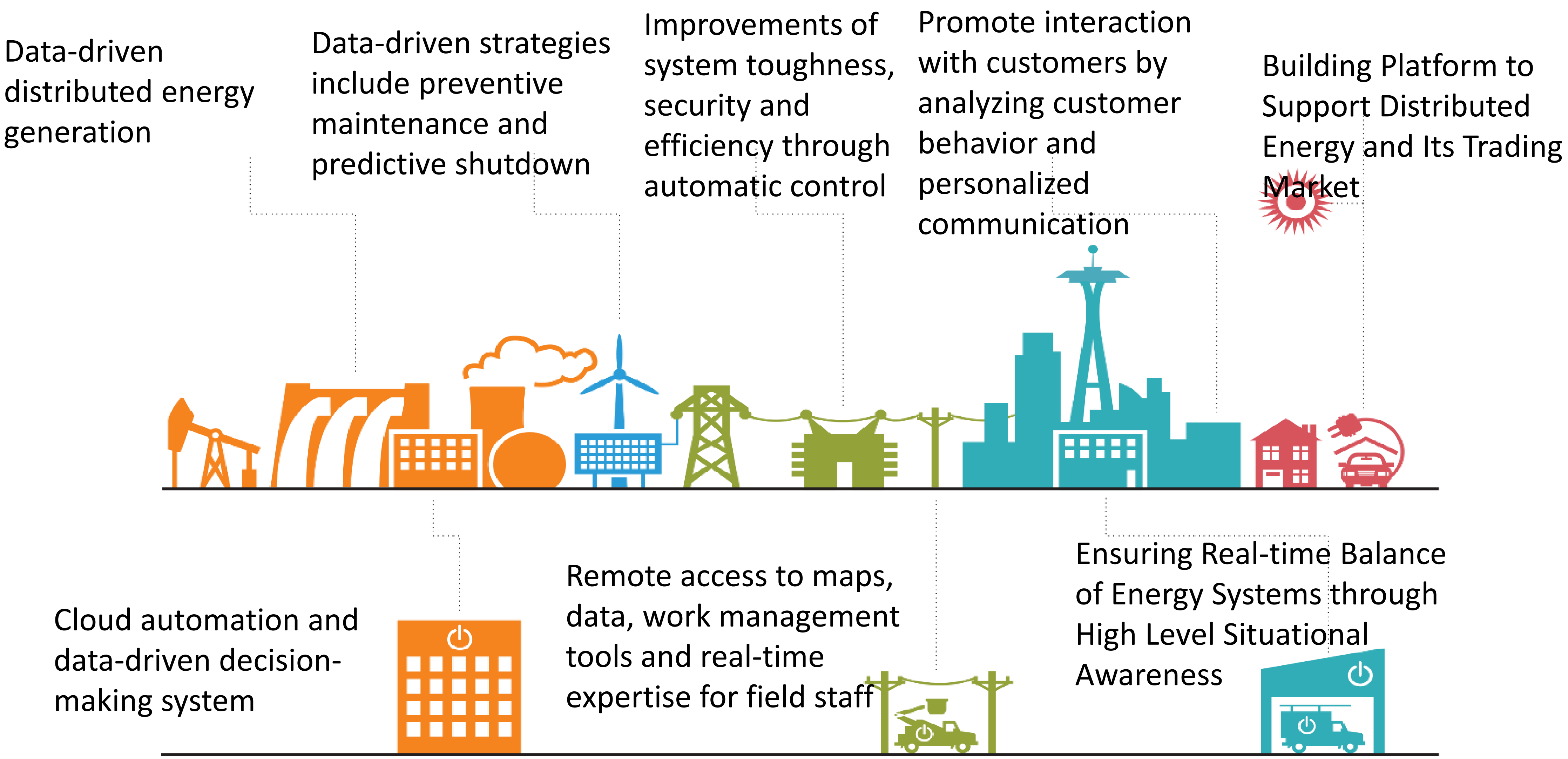
Data-driven Power Flow Linearization and Operation Mode Analysis

张宁 ningzhang@tsinghua.edu.cn

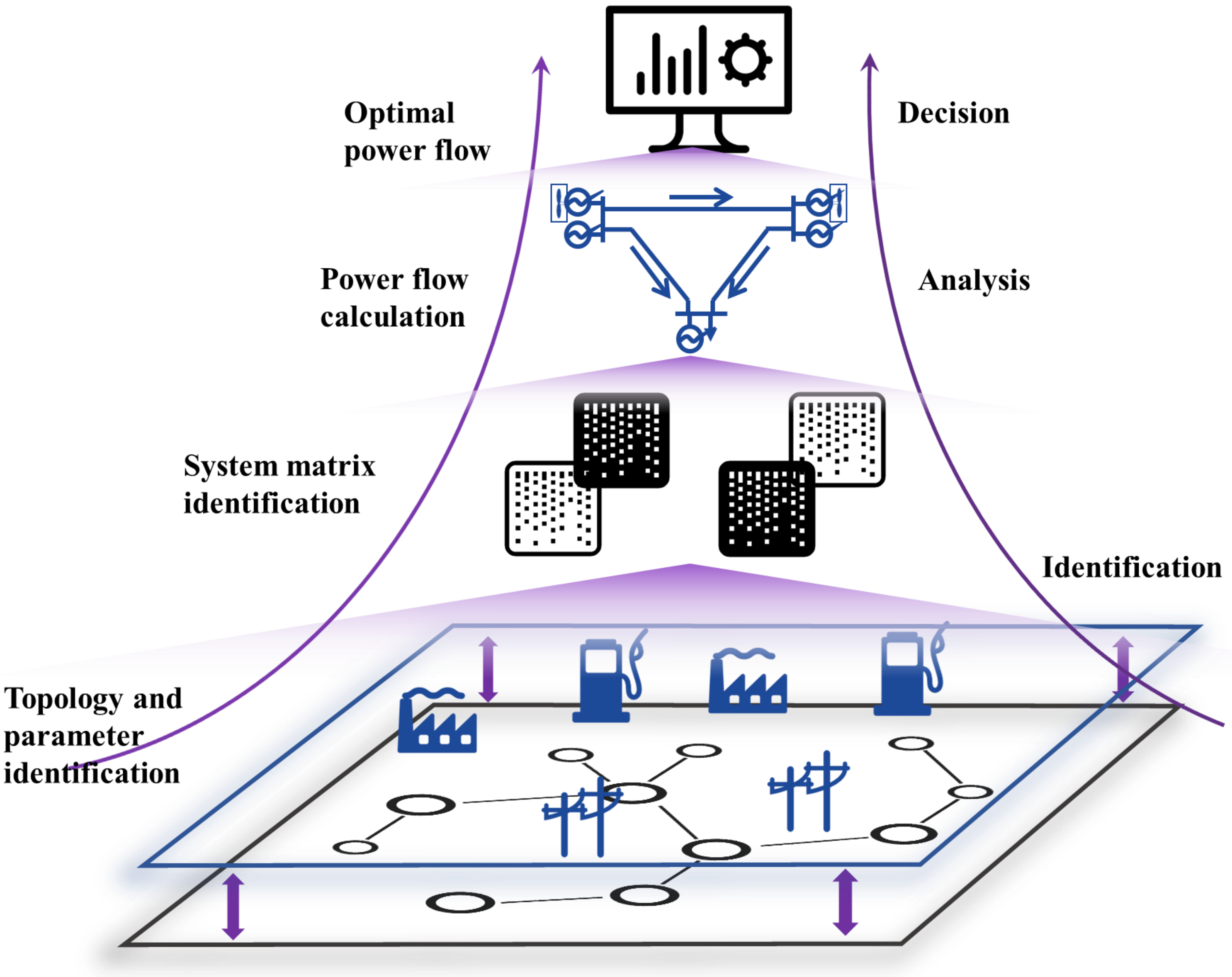
Contents

- **Big Data Applications in Power Systems**
- Data-driven Power Flow Linearization
- Data-driven Power System Operation Mode Analysis

Big Data Applications in Power Systems



Big Data Applications in Power Systems



Contents

- Big Data Applications in Power Systems
- **Data-driven Power Flow Linearization**
- Data-driven Power System Operation Mode Analysis

AC Power Flow Equations

Branch power flow equations.

$$\dot{S}_{ij} = \dot{V}_i \hat{\dot{I}}_{ij}$$

$$P_{ij} = (V_i^2 - V_i V_j \cos\theta_{ij}) g_{ij} - V_i V_j \sin\theta_{ij} b_{ij}$$

$$Q_{ij} = -(V_i^2 - V_i V_j \cos\theta_{ij}) b_{ij} - V_i V_j \sin\theta_{ij} g_{ij}$$

Power flow equations.

$$P_i = V_i \sum_{j \in i} V_j (G_{ij} \cos\theta_{ij} + B_{ij} \sin\theta_{ij}) \;\; i = 1, 2, \ldots, N$$

$$Q_i = V_i \sum_{j \in i} V_j (G_{ij} \sin\theta_{ij} - B_{ij} \cos\theta_{ij}) \;\; i = 1, 2, \ldots, N$$

DC Power Flow Equations

DC power flow equations are linear power flow equations that only describe the relationship between P and θ.

$$P_i = V_i \sum_{j \in i} V_j (G_{ij} \cos\theta_{ij} + B_{ij} \sin\theta_{ij}) \; i = 1, 2, ..., N$$

$$\theta_{ij} \approx 0, \; \cos\theta_{ij} = 1, \quad \sin\theta_{ij} = \theta_{ij}; \; r_{ij} << x_{ij}, \text{ignore } r_{ij}$$

$$P_{ij} = \frac{\theta_i - \theta_j}{\boxed{x_{ij}}}$$

Resistance of branch *ij*

Decoupled linearized power flow (DLPF)

$$\begin{bmatrix} P_{ij}/V_i \\ Q_{ij}/V_i \end{bmatrix} = -\begin{bmatrix} b_{ij} & -g_{ij} \\ g_{ij} & b_{ij} \end{bmatrix} \begin{bmatrix} V_j \sin \vartheta_{ij} \\ V_i - V_j \cos \vartheta_{ij} \end{bmatrix}$$

Approximation

$V_i \approx 1$, $V_j \approx 1$, $\cos \vartheta_{ij} \approx 1$

$$\begin{bmatrix} P_{ij} \\ Q_{ij} \end{bmatrix} = -\begin{bmatrix} b_{ij} & -g_{ij} \\ g_{ij} & b_{ij} \end{bmatrix} \begin{bmatrix} \vartheta_i - \vartheta_j \\ V_i - V_j \end{bmatrix}$$

Written in nodal injection form

$$\begin{bmatrix} P \\ Q \end{bmatrix} = -\begin{bmatrix} B' & -G \\ G' & B \end{bmatrix} \begin{bmatrix} \vartheta \\ V \end{bmatrix}$$

B', G' does not consider the shunt capacitor

对于V的幅值计算比较准确，无论在输电网还是在配电网中

Jingwei Yang, Ning Zhang, Chongqing Kang and Qing Xia. A State-Independent Linear Power Flow Model with Accurate Estimation of Voltage Magnitude, IEEE Transactions on Power Systems, 2017, 32(5): 3607-3617.

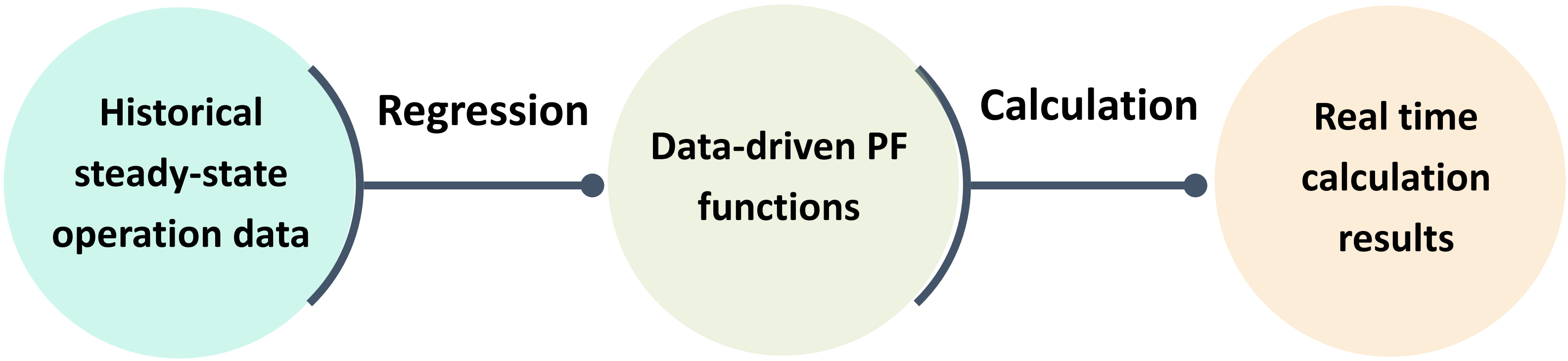
Data-driven Power Flow Linearization

Conventional linearization

- Linearize the nonlinear power flow equations (or the AC power flow equations) by some physical assumptions. For example, in the DC power flow:

$$\theta_{ij} \approx 0,\ \cos\theta_{ij} = 1, \quad \sin\theta_{ij} = \theta_{ij};\ r_{ij} << x_{ij}, \text{ignore } r_{ij}$$

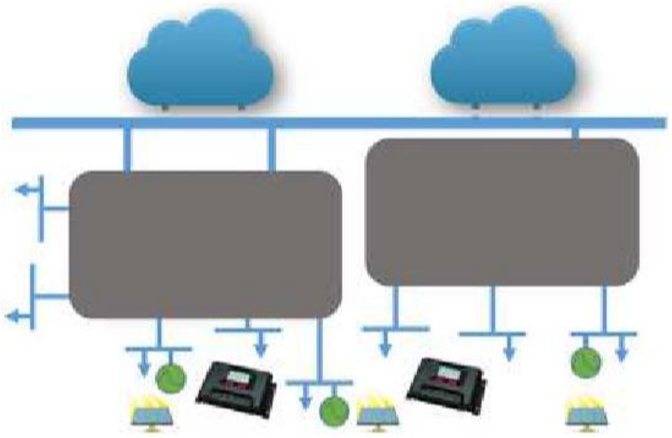
Data-driven linearization

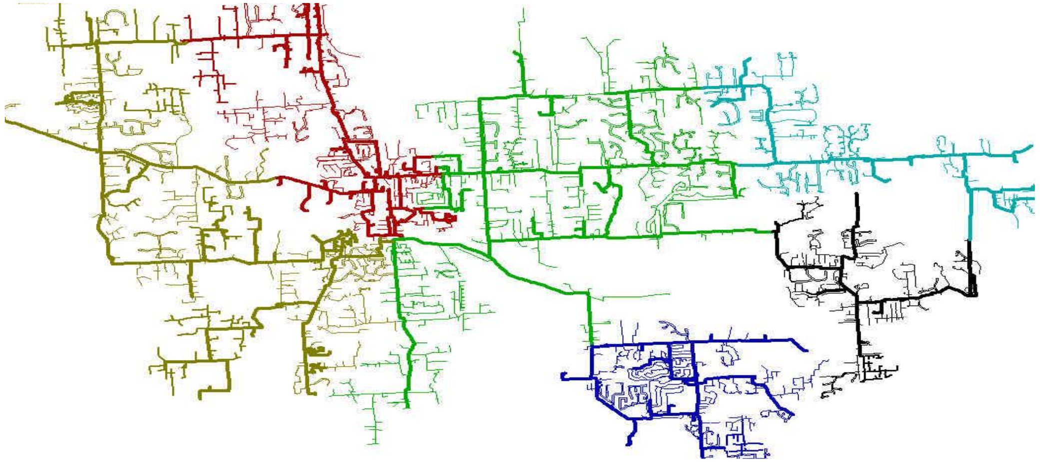


- Linearize the nonlinear power flow equations by the historical steady-state operation data

Data-driven Power Flow Linearization

Why data-driven?





- Do not require knowledge of the system topologies and parameters

The exact system topologies, element parameters, and the control logic of active control devices are difficult to model accurately in some distribution network.

- Improve the linearization accuracy of PF calculations

The measurement data reflects the operation status more efficiently than equivalent parameters. (e.g. parameters may change due to the atmospheric condition and aging)

Data-driven Power Flow Linearization

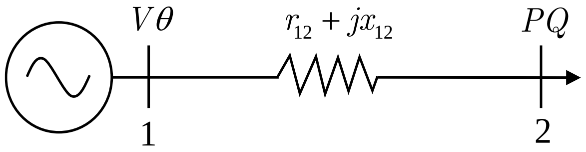
Why regression? Why not use neural network based, kernel based, or tree based methods to describe the power flow equations?

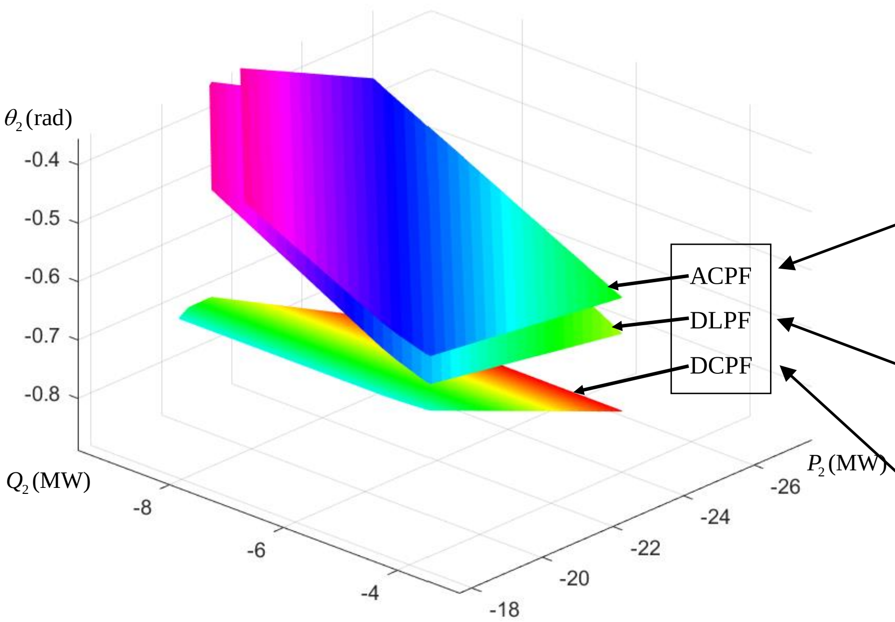
- Data-driven methods are more valuable in linearized applications rather than nonlinear calculations.
- Neural network based, kernel based, or tree based methods are not computational efficient in power system analysis.

Data-driven Power Flow Linearization

Why regression? Why not use neural network based, kernel based, or tree based methods to describe the power flow equations?

- The non-linear ACPF surface has a high degree of linearity, linear regression can have good performance, see the two-bus visualization:





$$P_i = V_i \sum_{j \in i} V_j (G_{ij} \cos\theta_{ij} + B_{ij} \sin\theta_{ij})$$

$$Q_i = V_i \sum_{j \in i} V_j (G_{ij} \sin\theta_{ij} - B_{ij} \cos\theta_{ij})$$

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

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

Data-driven Power Flow Linearization

Formulation:

(P, Q) as a function of (V, θ)

Regression

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

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

Unknown

Net loss

Unknown

Calculation

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

Why not (V, θ) as a function of (P, Q) ?

Y. Liu, N. Zhang, Y. Wang, J. Yang and C. Kang, "Data-Driven Power Flow Linearization: A Regression Approach," in IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 2569-2580, May 2019, doi: 10.1109/TSG.2018.2805169.

Data-driven Power Flow Linearization

Why not (V, θ) as a function of (P, Q) ?

- It can calculate PF when considering different bus types

Recall that in our aforementioned page:

PQ bus	Load buses, interconnect buses, generator buses without AVR
PV bus	Generator buses with AVR
$V\theta$ bus	Essential for calculation. Generator with sufficient regulating capability.

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

Unknown

$$\Downarrow$$

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

Reversible

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

If (V, θ) as a function of (P, Q) the related matrix is not reversible because of some zero injection interconnect buses

Data-driven Power Flow Linearization

Relationship with Physical Parameter Matrices:

Power system matrices

ACPF function

Jacobian matrix

$$\begin{bmatrix} \mathbf{B}\cos\boldsymbol{\theta} & -\mathbf{G}\cos\boldsymbol{\theta} \\ \mathbf{G}\cos\boldsymbol{\theta} & \mathbf{B}\cos\boldsymbol{\theta} \end{bmatrix} - \begin{bmatrix} \mathbf{G}\sin\boldsymbol{\theta} & \mathbf{B}\sin\boldsymbol{\theta} \\ -\mathbf{B}\sin\boldsymbol{\theta} & \mathbf{G}\sin\boldsymbol{\theta} \end{bmatrix} - \begin{bmatrix} -\mathbf{Q} & \mathbf{P} \\ \mathbf{P} & \mathbf{Q} \end{bmatrix}$$

Constant Jacobian matrix

$$\boxed{\begin{bmatrix} \mathbf{B} & -\mathbf{G} \\ \mathbf{G} & \mathbf{B} \end{bmatrix}}$$

DCPF function

X matrix

$$\boxed{\mathbf{B}^{-1}}$$

Regression matrices

Regression matrix of forward regression

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

Regression matrix of inverse regression

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

These relationships can serve as an indicator of overfitting

Data-driven Power Flow Linearization

Challenges of regression:

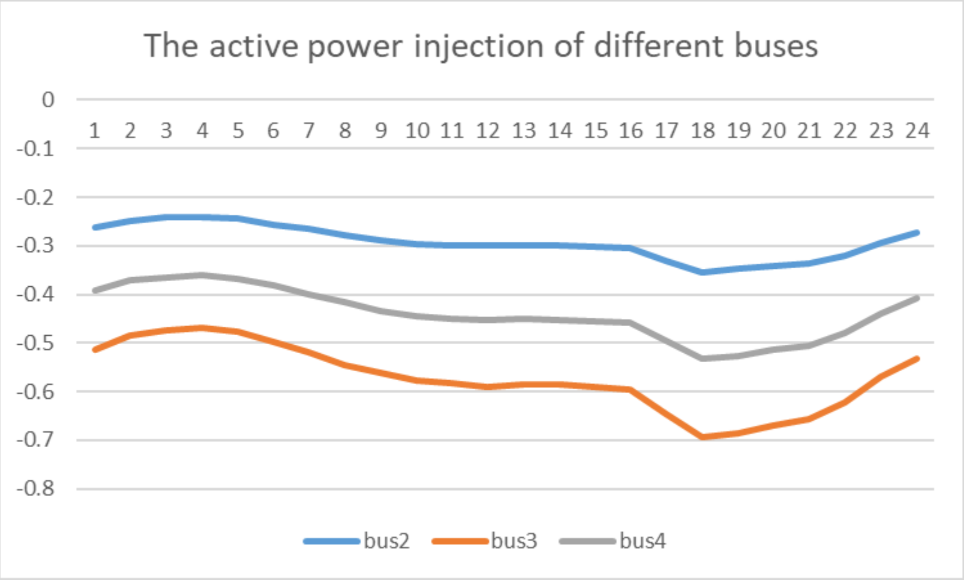
- to address the collinearity of data:
  - collinearity among the voltage angle and magnitude data is inevitable because of the similar rise and fall patterns among the different buses
  - result in ill-conditioned regression and larger errors of PF calculation

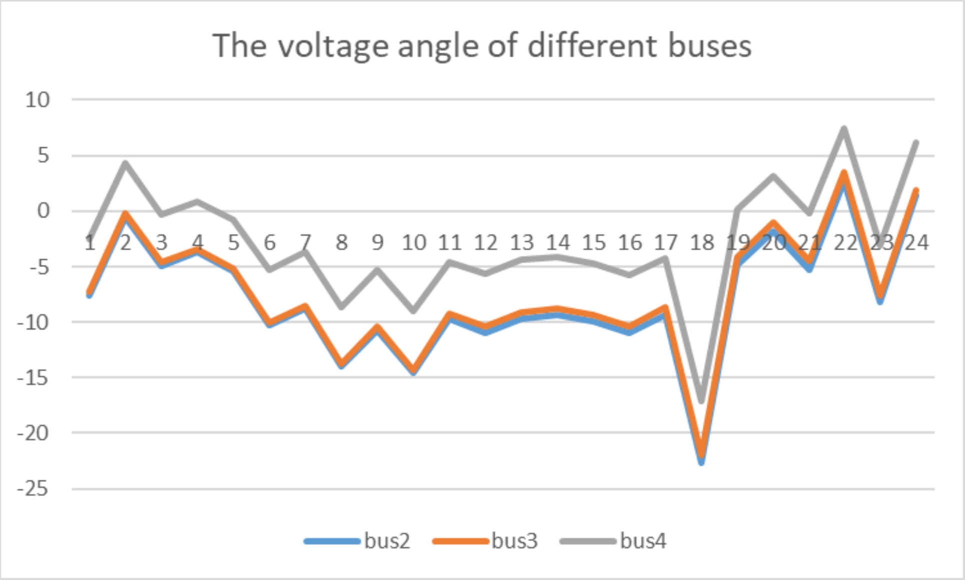
- to avoid overfitting:
  - the number of variables in the regression parameter matrices for large power systems may be far greater than the amount of historical operation data that represents the current system situation

- a Partial least square-based regression is proposed

Data-driven Power Flow Linearization

An visualization of collinearity:





A day (24 hours) profile of active power injections and voltage angles of NREL-118 system.

Data-driven Power Flow Linearization

Why not ordinary least squares?

Recall that for ordinary least squares:

$$\mathbf{Y} = \mathbf{A}\mathbf{X}$$
$$\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

When there is collinearity:

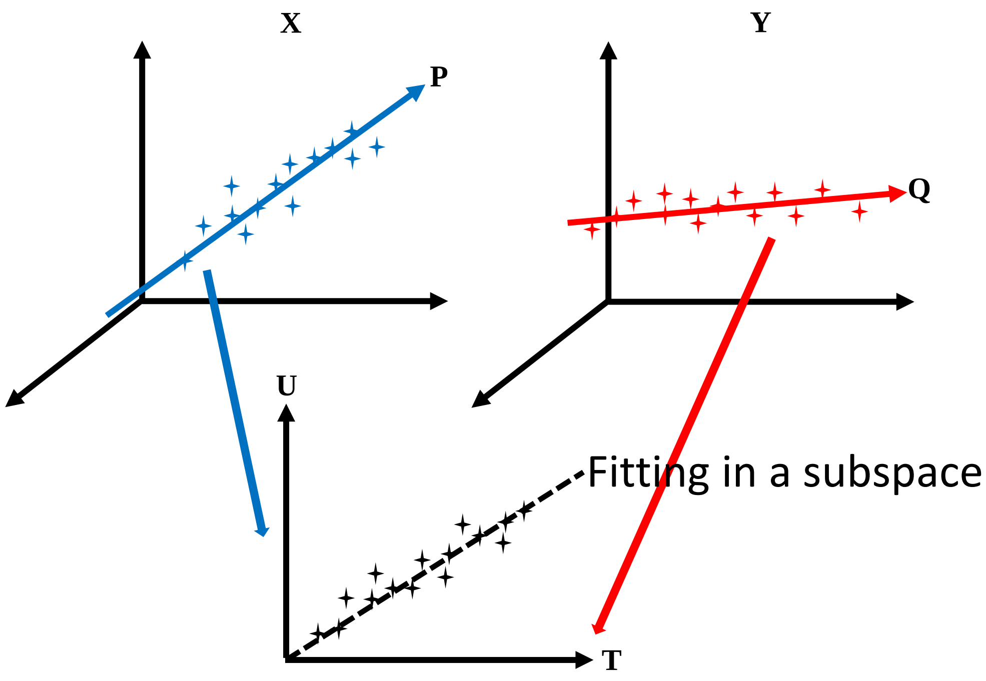
Small

$$\lambda_0 + \lambda_1 X_{1i} + \lambda_2 X_{2i} + \cdots + \lambda_k X_{ki} + \boxed{v_i} = 0.$$

In this case, the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ has an inverse, but is ill-conditioned so that a given computer algorithm may or may not be able to compute an approximate inverse

Data-driven Power Flow Linearization

Partial least square:



PLS can address the collinearity and lack of observations by projecting the predicted variables and the observable variables to a new space

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}$$

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

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 $\mathbf{X}$.

Data-driven Power Flow Linearization

Partial least square:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}$$

$\mathbf{P}, \mathbf{Q}$ ➡ Loading matrices, *p* extracted components of **X** and **Y**

$\mathbf{T}, \mathbf{U}$ ➡ Score matrices, projections of **X**, **Y** on **P**, **Q**

To find the principle components of **X** and **Y**
WHILE maintaining the greatest correlation ➡ Maximize the covariance of **T** and **U**

NIPALS: Nonlinear Iterative Partial Least Squares algorithm

Objective: To find weight vectors **w, c** such that

$$(\mathbf{w}, \mathbf{c}) = \arg\max_{|\mathbf{r}|=|\mathbf{s}|=1}[\mathrm{cov}(\mathbf{Xr}, \mathbf{Ys})]^2$$

➡ $$[\mathrm{cov}(\mathbf{t}, \mathbf{u})]^2 = [\mathrm{cov}(\mathbf{Xw}, \mathbf{Yc})]^2 = \max_{|\mathbf{r}|=|\mathbf{s}|=1}[\mathrm{cov}(\mathbf{Xr}, \mathbf{Ys})]^2$$

[1] R. Rosipal and N. Krämer, “Overview and recent advances in partial least squares,” in *Proc. Int. Conf. Subspace Latent Struct. Feature Selection*, 2005, pp. 34–51.
[2] https://zhuanlan.zhihu.com/p/414061371

Data-driven Power Flow Linearization

NIPALS: Nonlinear Iterative Partial Least Squares algorithm

Objective: To find weight vectors **w, c** such that

$$(\mathbf{w},\mathbf{c}) = \arg\max_{|\mathbf{r}|=|\mathbf{s}|=1}[\mathrm{cov}(\mathbf{Xr},\mathbf{Ys})]^2$$

$$\Rightarrow \quad [\mathrm{cov}(\mathbf{t},\mathbf{u})]^2 = [\mathrm{cov}(\mathbf{Xw},\mathbf{Yc})]^2 = \max_{|\mathbf{r}|=|\mathbf{s}|=1}[\mathrm{cov}(\mathbf{Xr},\mathbf{Ys})]^2$$

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F}$$

Initialization: ***u* := *y*$_j$** for some *j*

iterations

$$\begin{aligned}
&1)\ \mathbf{w} = \mathbf{X}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})\\
&2)\ \|\mathbf{w}\| \rightarrow 1\\
&3)\ \mathbf{t} = \mathbf{Xw}\\
&4)\ \mathbf{c} = \mathbf{Y}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})\\
&5)\ \|\mathbf{c}\| \rightarrow 1\\
&6)\ \mathbf{u} = \mathbf{Yc}
\end{aligned}$$

The iteration equals to solving the weight vector **w** for:

$$\mathbf{X}^\mathrm{T}\mathbf{Y}\mathbf{Y}^\mathrm{T}\mathbf{X}\mathbf{w} = \lambda\mathbf{w}$$

Which is the principal component of $\mathbf{X}^\mathrm{T}\mathbf{Y}$

[1] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Proc. Int. Conf. Subspace Latent Struct. Feature Selection*, 2005, pp. 34–51.
[2] https://zhuanlan.zhihu.com/p/414061371

Data-driven Power Flow Linearization

NIPALS: Nonlinear Iterative Partial Least Squares algorithm

Objective: To find weight vectors **w, c** such that

$$(\mathbf{w},\mathbf{c}) = \arg\max_{|\mathbf{r}|=|\mathbf{s}|=1}[\mathrm{cov}(\mathbf{Xr},\mathbf{Ys})]^2$$

$$[\mathrm{cov}(\mathbf{t},\mathbf{u})]^2 = [\mathrm{cov}(\mathbf{Xw},\mathbf{Yc})]^2 = \max_{|\mathbf{r}|=|\mathbf{s}|=1}[\mathrm{cov}(\mathbf{Xr},\mathbf{Ys})]^2$$

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F}$$

Initialization: $\boldsymbol{u} := \boldsymbol{y}_j$ for some j

iterations

$$
\begin{aligned}
&1)\ \mathbf{w} = \mathbf{X}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})\\
&2)\ \|\mathbf{w}\| \to 1\\
&3)\ \mathbf{t} = \mathbf{Xw}\\
&4)\ \mathbf{c} = \mathbf{Y}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})\\
&5)\ \|\mathbf{c}\| \to 1\\
&6)\ \mathbf{u} = \mathbf{Yc}
\end{aligned}
$$

Score vectors:

$$\mathbf{t} = \mathbf{Xw}$$
$$\mathbf{u} = \mathbf{Yc}$$

Loading vectors :

$$\mathbf{p} = \mathbf{X}^T\mathbf{t}/\left(\mathbf{t}^T\mathbf{t}\right)$$
$$\mathbf{q} = \mathbf{Y}^T\mathbf{u}/\left(\mathbf{u}^T\mathbf{u}\right)$$

Update **X**, **Y:**

$$\mathbf{X} = \mathbf{X} - \mathbf{tp}^T$$
$$\mathbf{Y} = \mathbf{Y} - \mathbf{uq}^T$$

Calculate the next
t, u, p, q

[1] R. Rosipal and N. Krämer, “Overview and recent advances in partial least squares,” in *Proc. Int. Conf. Subspace Latent Struct. Feature Selection*, 2005, pp. 34–51.

[2] https://zhuanlan.zhihu.com/p/414061371

Numerical results

Data generation:

◻ Monte Carlo simulation:

- meshed transmission grids: IEEE 5, 30, 57, and 118-bus systems
- radial distribution grids: IEEE 33-bus system, the modified 123-bus system

◻ Public testing data:

- the NREL-118 test system (data collinearity)

Numerical results

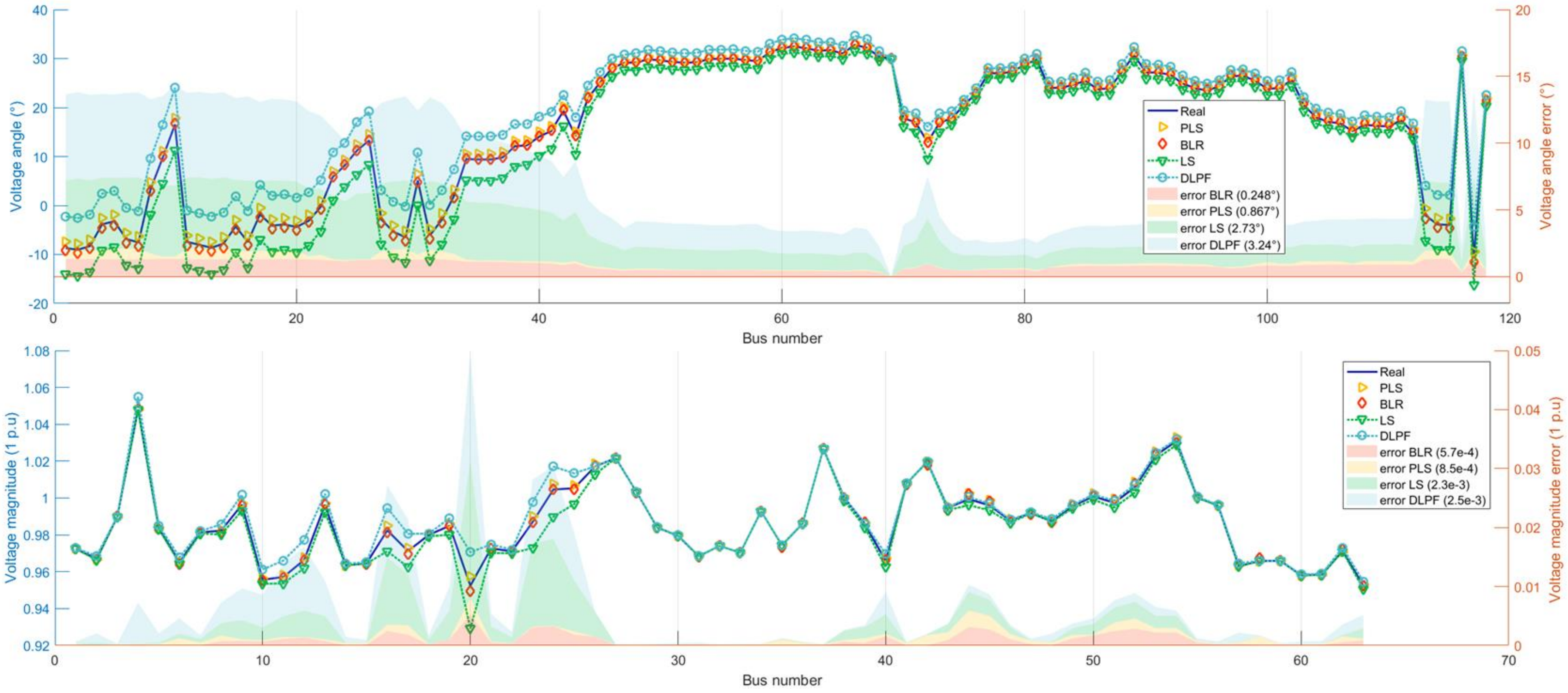
Basic results:

Cases	Size of training data	Size of testing data	Forward calculation				Inverse calculation		
			Errors	DCPF	DLPF	PLS	Errors	DLPF	PLS
IEEE 5	100	300	P	24.11	1.117	**0.412**	θ	0.020	**8.2e-4**
			Q	---	66.21	**0.940**	V	7.8e-4	**2.0e-5**
IEEE 30	100	300	P	12.49	0.578	**0.034**	θ	0.154	**1.9e-3**
			Q	---	12.66	**0.404**	V	9.9e-4	**1.0e-5**
IEEE 33	100	300	P	67.05	1.114	**0.012**	θ	0.028	**4.3e-4**
			Q	---	0.759	0.044	V	2.0e-3	**7.3e-6**
IEEE 57	300	300	P	98.11	7.343	**0.262**	θ	0.215	**0.036**
			Q	---	26.83	**0.300**	V	7.1e-3	**2.1e-4**
IEEE 118	300	300	P	16.89	4.546	**0.061**	θ	2.593	**0.074**
			Q	---	77.85	**1.096**	V	1.9e-3	**1.2e-4**
NREL 118	300	300	P	85.90	9.486	**0.161**	θ	3.003	0.622
			Q	---	107.4	**0.486**	V	2.3e-3	**6.3e-4**
Modified 123	300	300	P	12.49	0.512	**0.007**	θ	0.091	**3.2e-4**
			Q	---	2.071	**0.003**	V	2.3e-3	**3.2e-6**

Numerical results

Calculation results under data collinearity:

- the NREL-118 test system

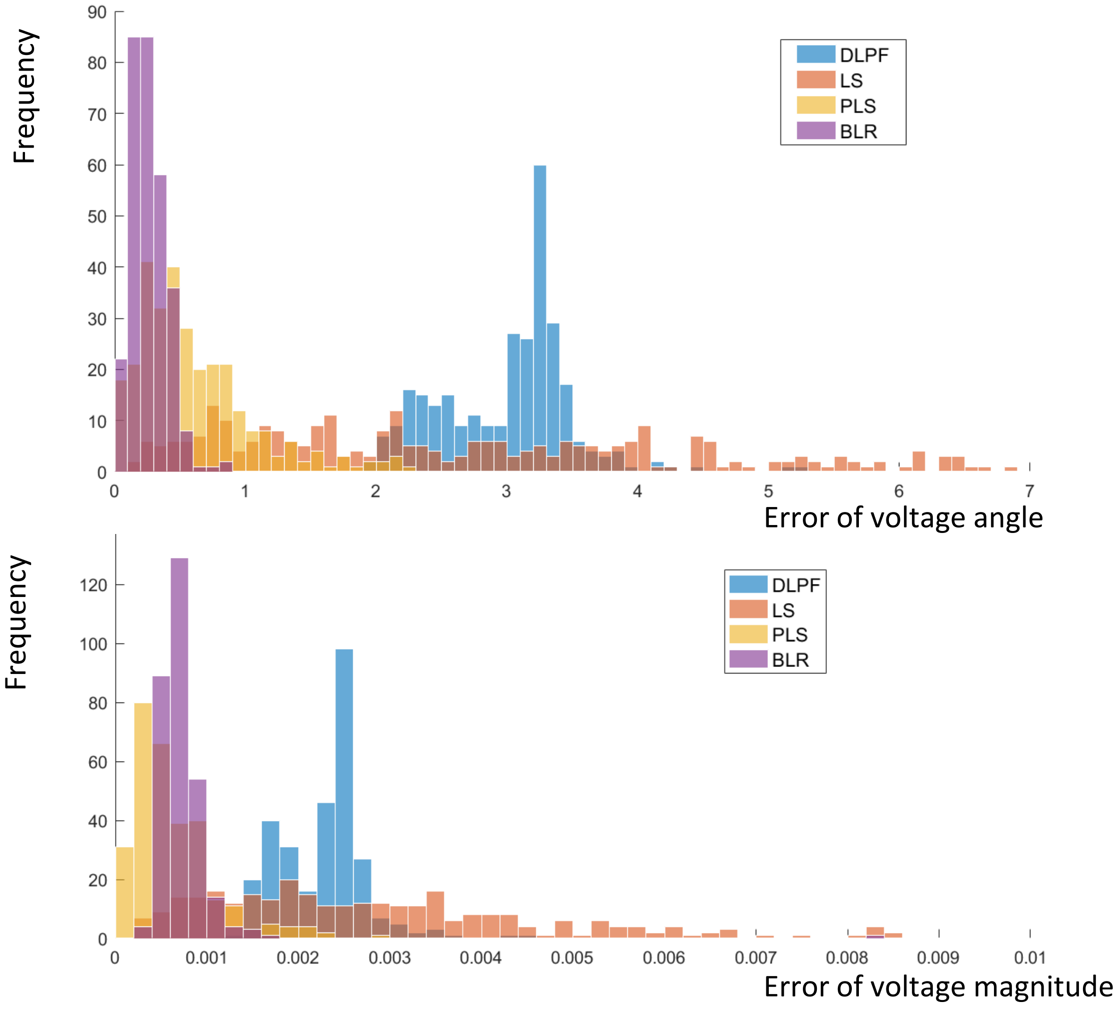


To show the robustness of the algorithm, the error in the figure is the largest among all groups in the NREL-118 test system.

Numerical results

Calculation results under data collinearity:

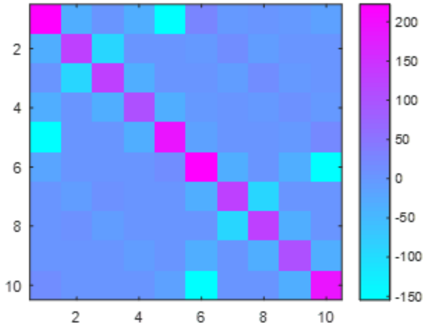
- the NREL-118 test system



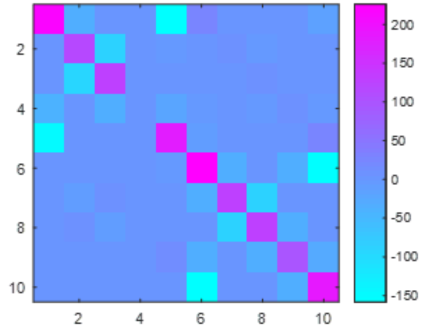
Numerical results

Regression Parameters：

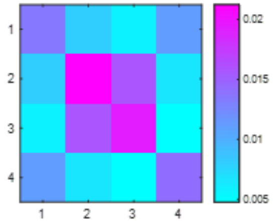
- IEEE 5-bus systems



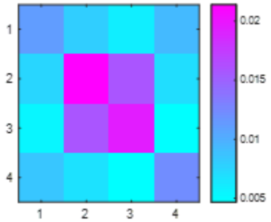
Constant Jacobian matrix



Forward regression parameter matrix under the PLS-based algorithm



The inverse matrix of **B** in DCPF



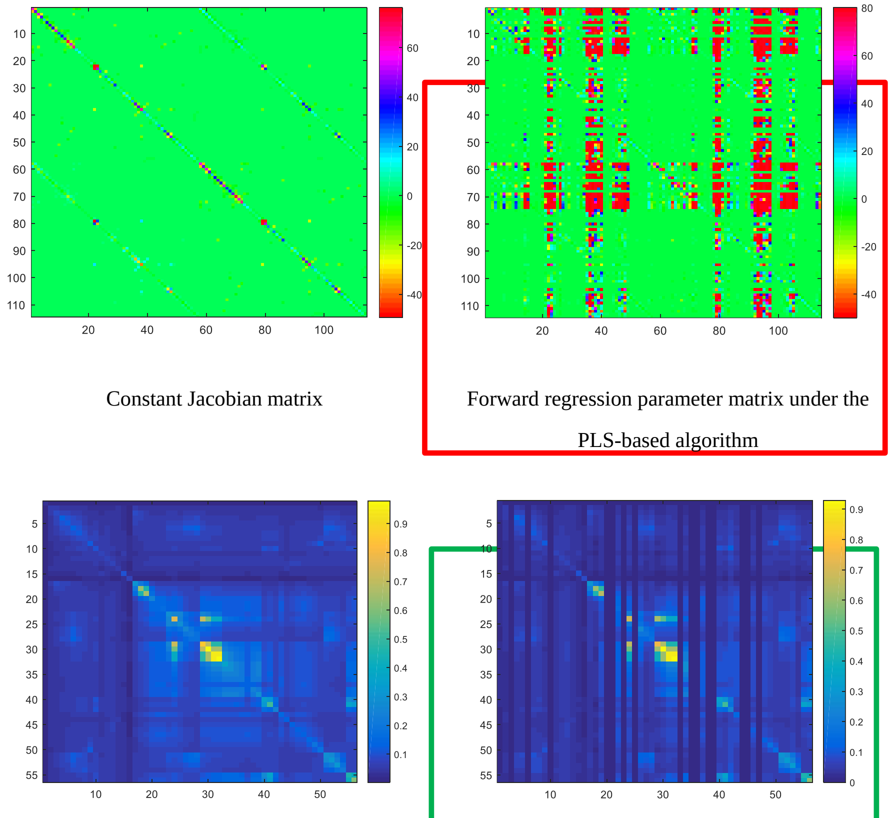
Inverse regression parameter matrix under the PLS-based algorithm

Comparisons between regression parameter matrices and several power system matrices of IEEE 5-bus system

Numerical results

Regression Parameters：

- IEEE 57-bus systems



Constant Jacobian matrix

Forward regression parameter matrix under the PLS-based algorithm

The inverse matrix of **B** in DCPF

Inverse regression parameter matrix under the PLS-based algorithm

Fig. A2. Comparisons between regression parameter matrices and several power system matrices of IEEE 57-bus system

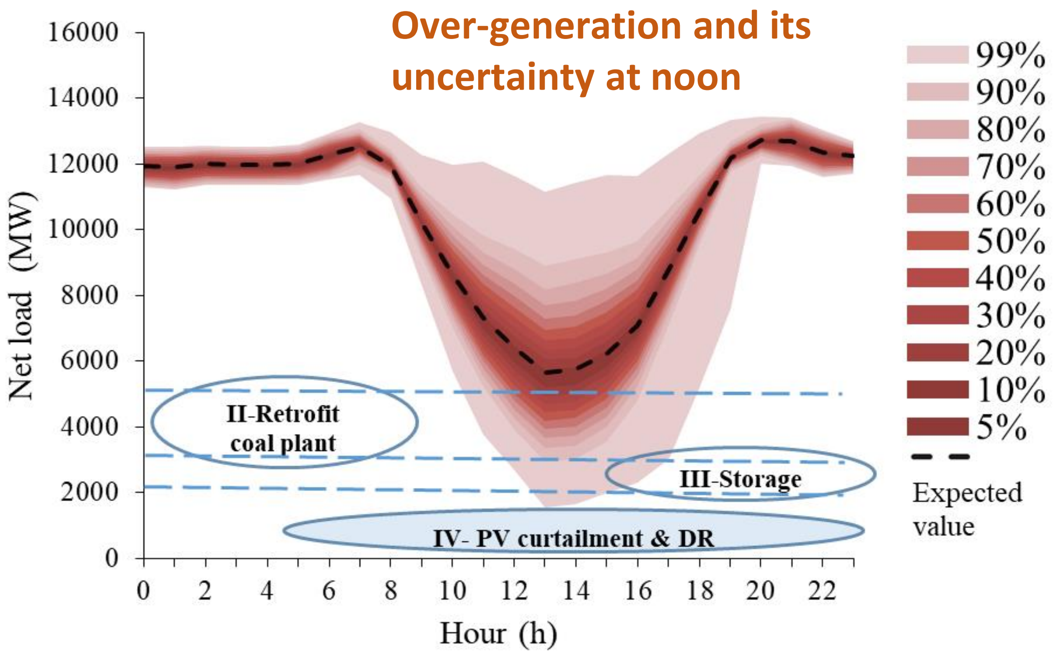
Contents

- Big Data Applications in Power Systems
- Data-driven Power Flow Linearization
- **Data-driven Power System Operation Mode Analysis**

Motivations

- The introduction of highly penetrated renewable energy will make the power system operation mode highly diversified and variable. These modes may not follow traditional empirical patterns.
- **How will the operation mode change with increasing renewable energy penetration ?**
- **How to pick typical/representative days in power system planning/operation analysis ?**

Probabilistic Duck Curve in Qinghai in 2020



Power system operation mode

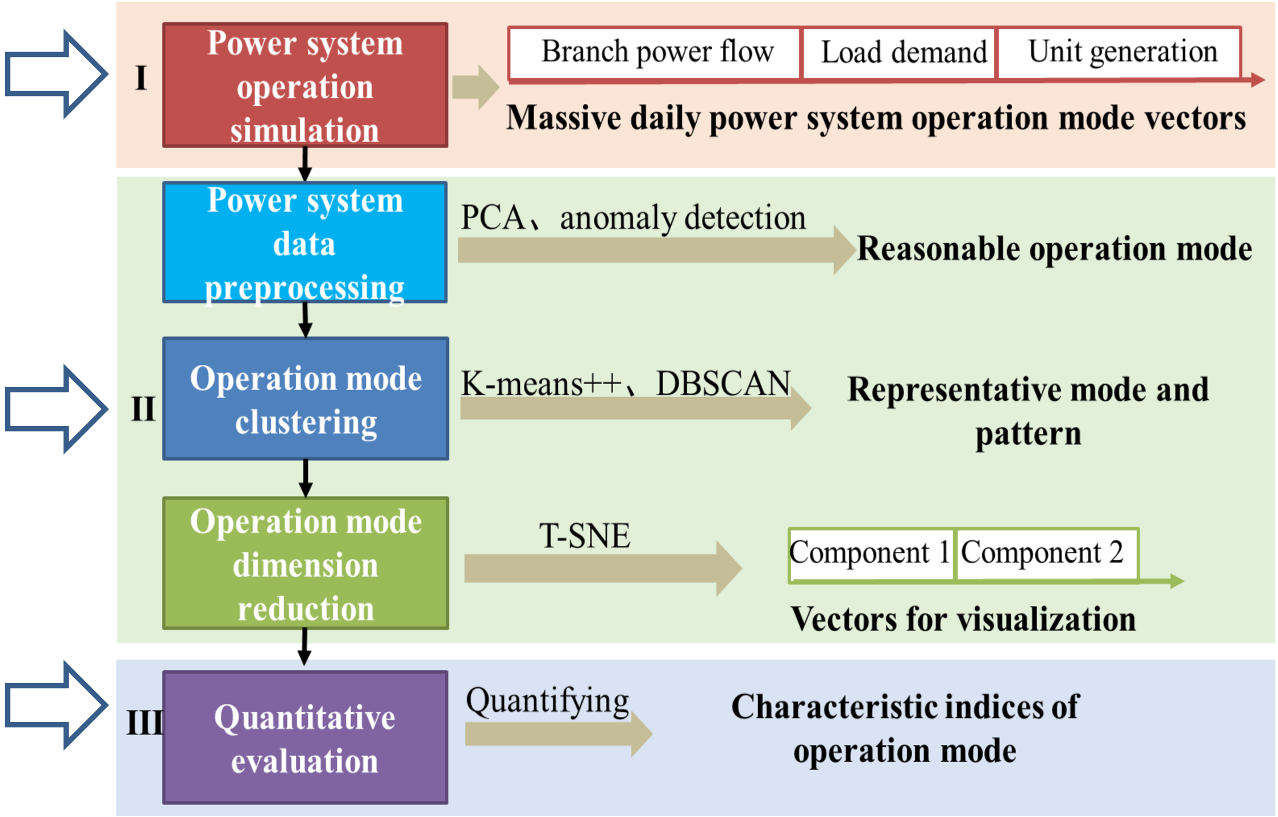
- **The power system operation mode definition:** the status of power system operation, which is determined by the generator outputs, load demand, transmission topology, and accordant power flow in a certain period, such as a day, an hour or a snapshot.

- **Identifying the power system operation mode pattern is a typical big-data analytic problem.**
  - These data are inherently high-dimensional and complexly coupled to one another
  - Those operation data will have a significant variation with time, which makes it very hard to find the patterns in large amounts of data.

Data-driven Framework

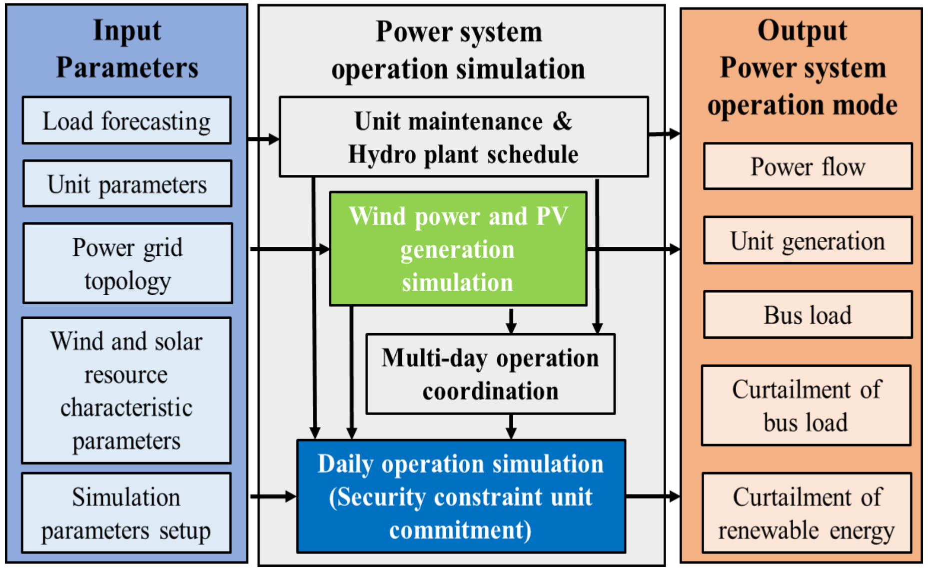
Three issues:

- **Power system operation data acquisition**: Sufficient operation data including power flow, load, and unit output are needed to form a complete year-round operation modes data set.
- **Pattern identification:** How to recognize the key characteristics of massive operation modes and identify the typical patterns and the number of patterns.
- **Visualization and evaluation:** How to visualize the high-dimensional operation mode data to provide intuitive understanding and quantify their characteristics.



Operation data acquisition

- Power system chronological operation simulation



An operation mode should include three aspects: energy source side, grid side, and load demand side.

$$\mathbf{p} = (\mathbf{g}_{1\times(|g|\times T_s)}, \mathbf{r}_{1\times(|r|\times T_s)}, \mathbf{f}_{1\times(|f|\times T_s)}, \mathbf{d}_{1\times(|d|\times T_s)}, \mathbf{d}^{\mathbf{c}}{}_{1\times(|d|\times T_s)})^T$$

generation schedule | renewable energy | power flow | load demand | load shedding

- Operation data can also be obtained through SCADA system

Principle to select algorithm

The selected algorithm must be suitable for high-dimensional and correlated data analysis.

- The selected preprocessing algorithm must be efficient with large amount of data;
- The clustering algorithm must be able to find complicate power system patterns under high renewable energy penetration;
- The dimension reduction and visualization algorithm should be able to decouple the correlation among high-dimensional features and map them into 2D/3D space for visualization and intuitive understanding.

Preprocessing using PCA

Linear Transformation $$\mathbf{P}' = \mathbf{H}^{\mathrm{T}}\mathbf{P}$$

Transform matrix H is composed by K principal vectors

Determining the number of principal components $$K = \arg_K \left(\sum_{i=1}^{K} \lambda_i \Big/ \sum_{i=1}^{N} \lambda_i \geq (1-\theta_0) \right)$$

Scenarios in Qinghai	**2017**	**2020**	**2025**
The number of original dimensions	>7000	>7000	>7000
The number of principal components	94	104	119

Dimension Reduction for Visualization

- The t-SNE algorithm is able to locally map the high-dimensional operation mode to a low-dimensional space while maintaining the global distribution of the operation mode.

Gaussian distribution

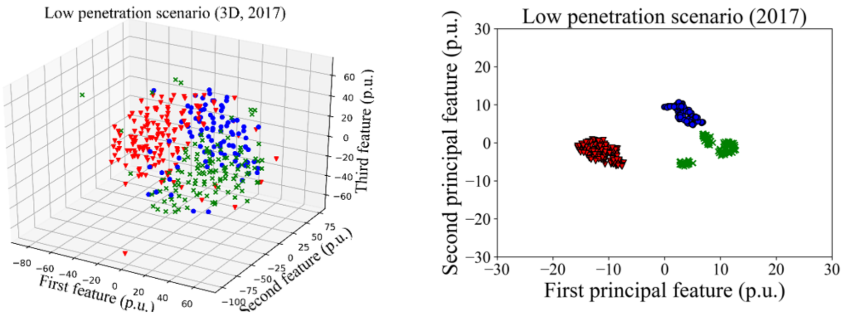
$$s_{ij} = \frac{\dfrac{\exp\left(-\dfrac{\|\mathbf{p}'_j - \mathbf{p}'_i\|^2}{2\sigma_j^2}\right)}{\sum_{k \neq j} \exp\left(-\dfrac{\|\mathbf{p}'_j - \mathbf{p}'_k\|^2}{2\sigma_j^2}\right)} + \dfrac{\exp\left(-\dfrac{\|\mathbf{p}'_i - \mathbf{p}'_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\dfrac{\|\mathbf{p}'_i - \mathbf{p}'_k\|^2}{2\sigma_i^2}\right)}}{2M}$$

minimize the difference

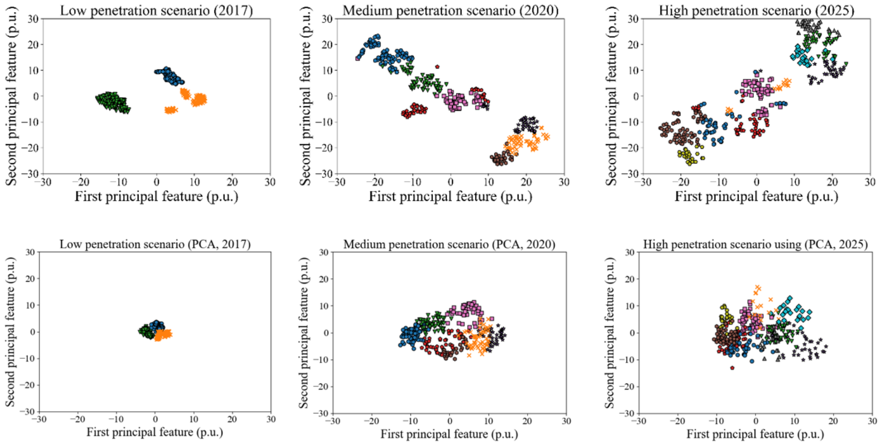
t-distribution

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{m \neq n} (1 + \|\mathbf{y}_m - \mathbf{y}_n\|^2)^{-1}}$$

2D vs 3D



PCA vs t-SNE



Quantitative Evaluation

A. Space Dispersion

High dimensional variance (HDV) is defined as the variance of high dimensional operation mode data after PCA:

$$HDV = \frac{1}{M} Tr(P'P'^{T})$$

B. Time variation

Operation mode switching frequency (OMSF) is the times that the operation mode pattern changes during a year

$$OMSF = \sum_{i=1}^{M} I(\mathbf{p}_i')$$

$$I(\mathbf{p}_i') = \begin{cases} 1 & if \ \mathbf{p}_i' \in \Omega \text{ and } \mathbf{p}_{i+1}' \notin \Omega \\ 0 & if \ \mathbf{p}_i' \in \Omega \text{ and } \mathbf{p}_{i+1}' \in \Omega \end{cases}$$

C. Seasonal consistency

Seasonal consistency (SC) is defined as the ratio of the number of days in which daily operation modes are unchanged to the total number of days in a certain season:

$$SC_j = \frac{M_j - \sum_{i=1}^{M_j} I(\mathbf{p}_i')}{M_j}$$

$$j \in \{\text{spring,summer,autumn,winter}\}$$

Average seasonal consistency (ASC)

$$ASC = \frac{1}{4} \sum_{j} SC_j$$

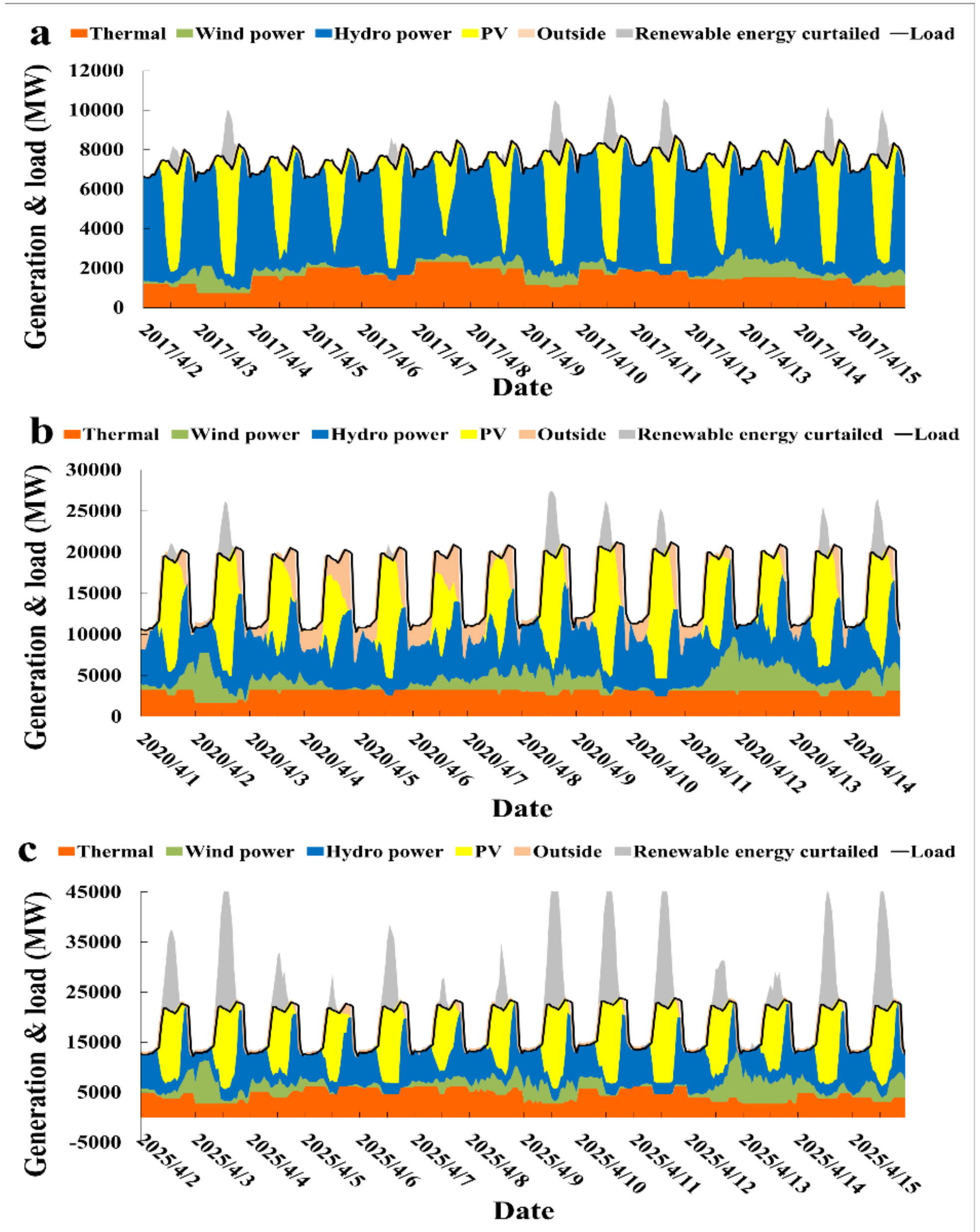
$$j \in \{\text{spring,summer,autumn,winter}\}$$

Case Study

- **Three renewable energy penetration scenarios are compared:**
- low penetration (20%) scenarios in 2017, medium penetration (33%) scenarios in 2020, high penetration (40%) scenarios in 2025.

Year	2017	2020	2025
Hydro-power (MW)	1169	1637	1900
Thermal Power(MW)	360	510	850
Wind Power (MW)	162	700	1081
PV (MW)	790	2000	3800
PV and Wind Power Capacity/ Total Capacity (%)	38	56	64
Total Load (GWh)	88000	141300	161300
Maximal Load (MW)	10000	22000	25000
PV and Wind Power Generation/Total Load (%)	**20**	**33**	**40**

Power system simulation results



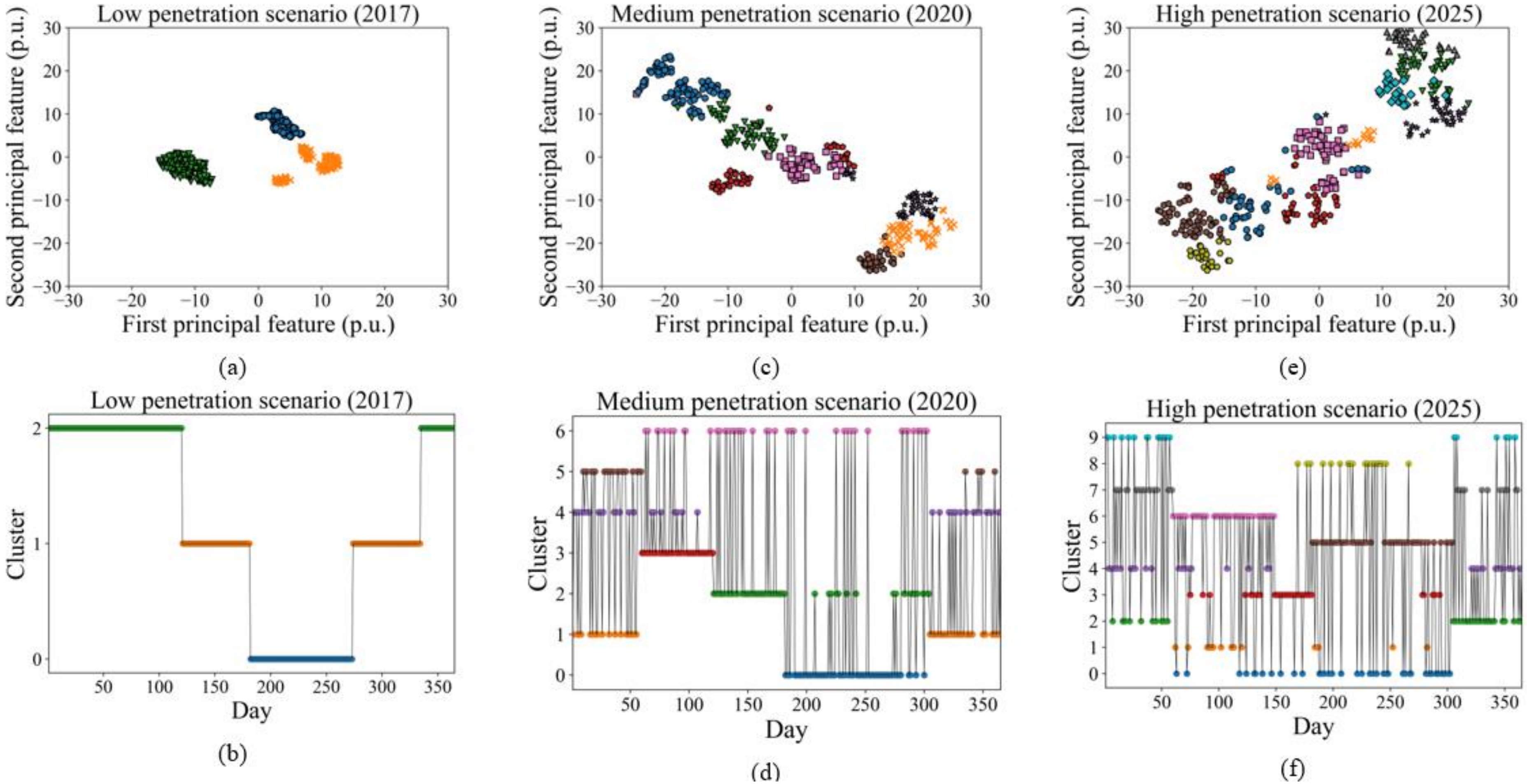
20%
Hydro-dominated

33%
PV and wind play an important role

40%
Intermittent renewable energy-dominated

Visualization of data-driven results

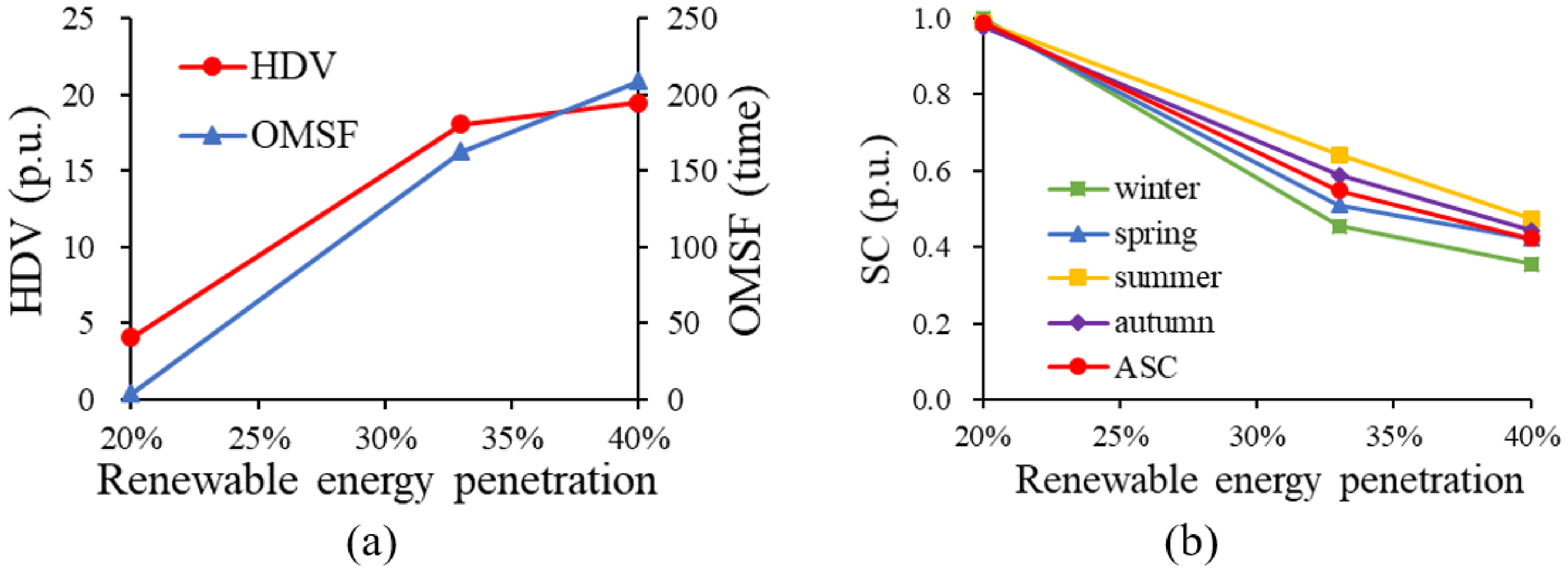
- Increasing space dispersion
- Increasing time variation
- Increasing pattern numbers
- Increasing patterns in each season



Quantification of data-driven results

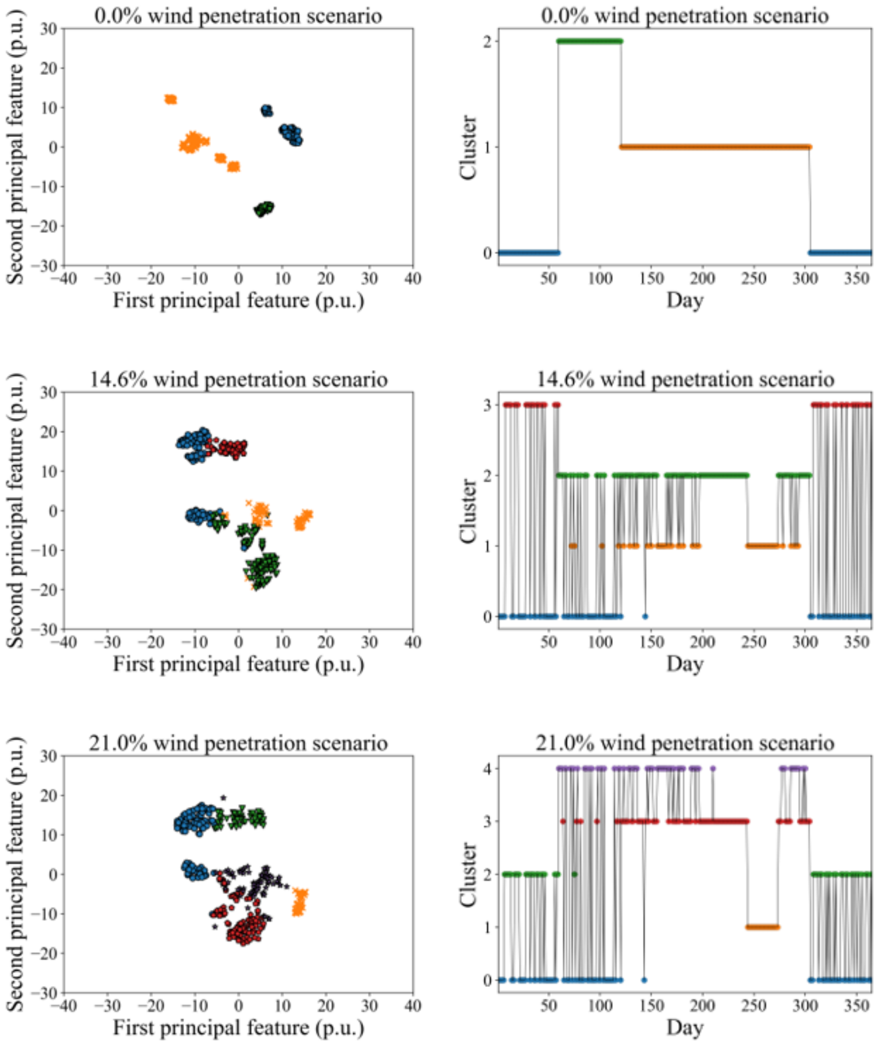
High dimensional variance (HDV) & Operation mode switching frequency

Seasonal consistency (SC)

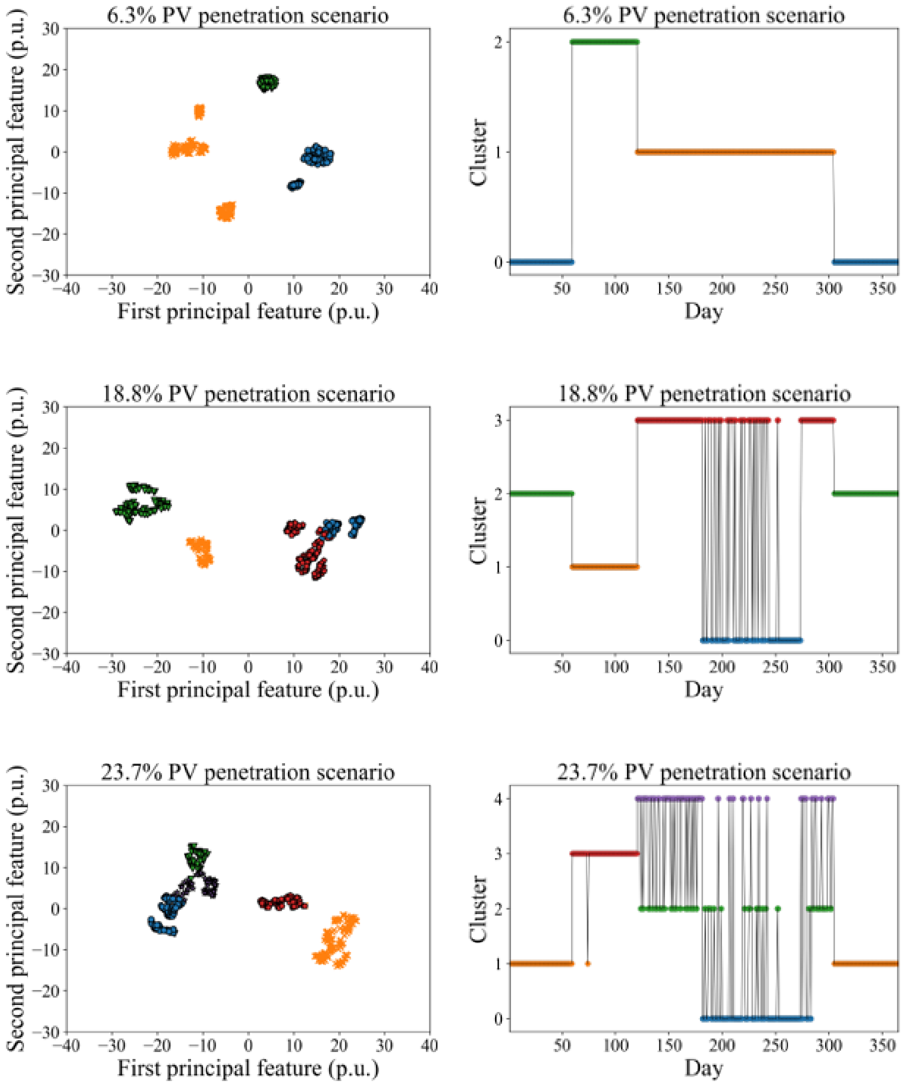


More results

Adding more wind power

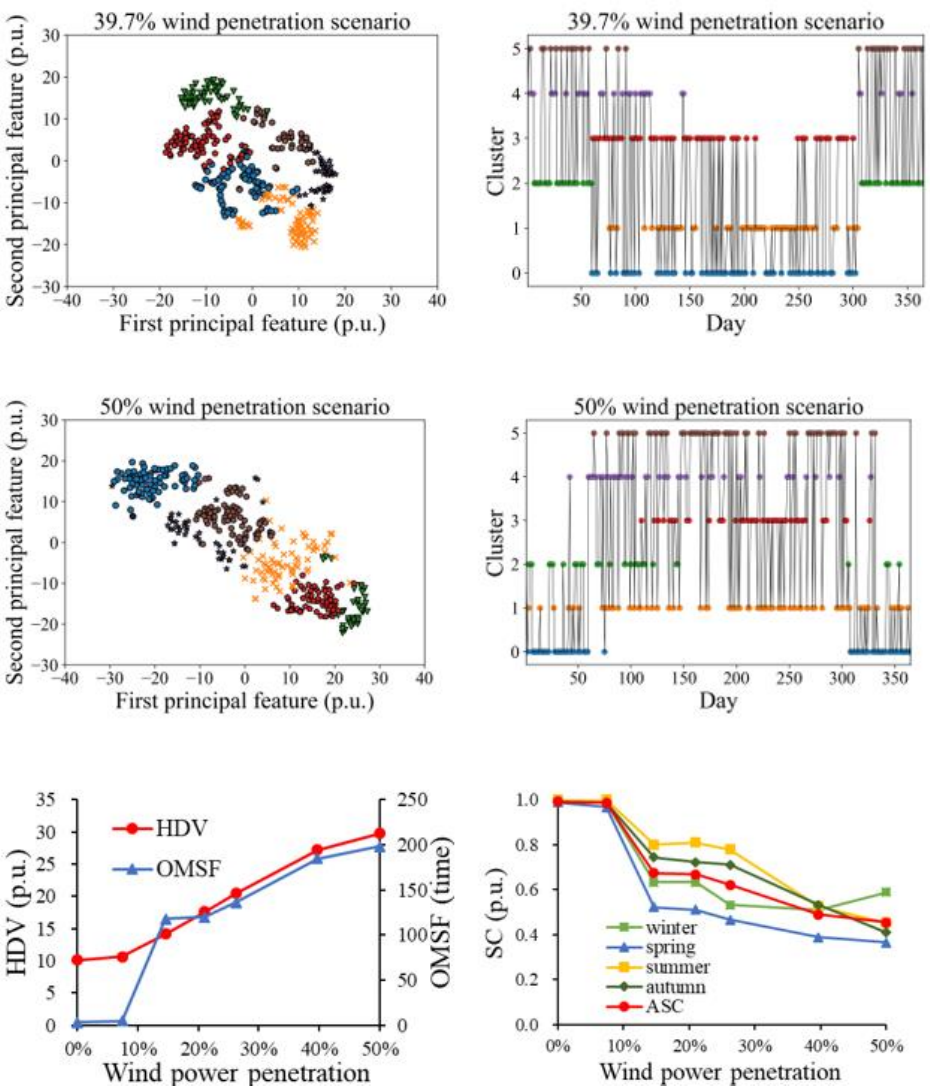


Adding more PV power

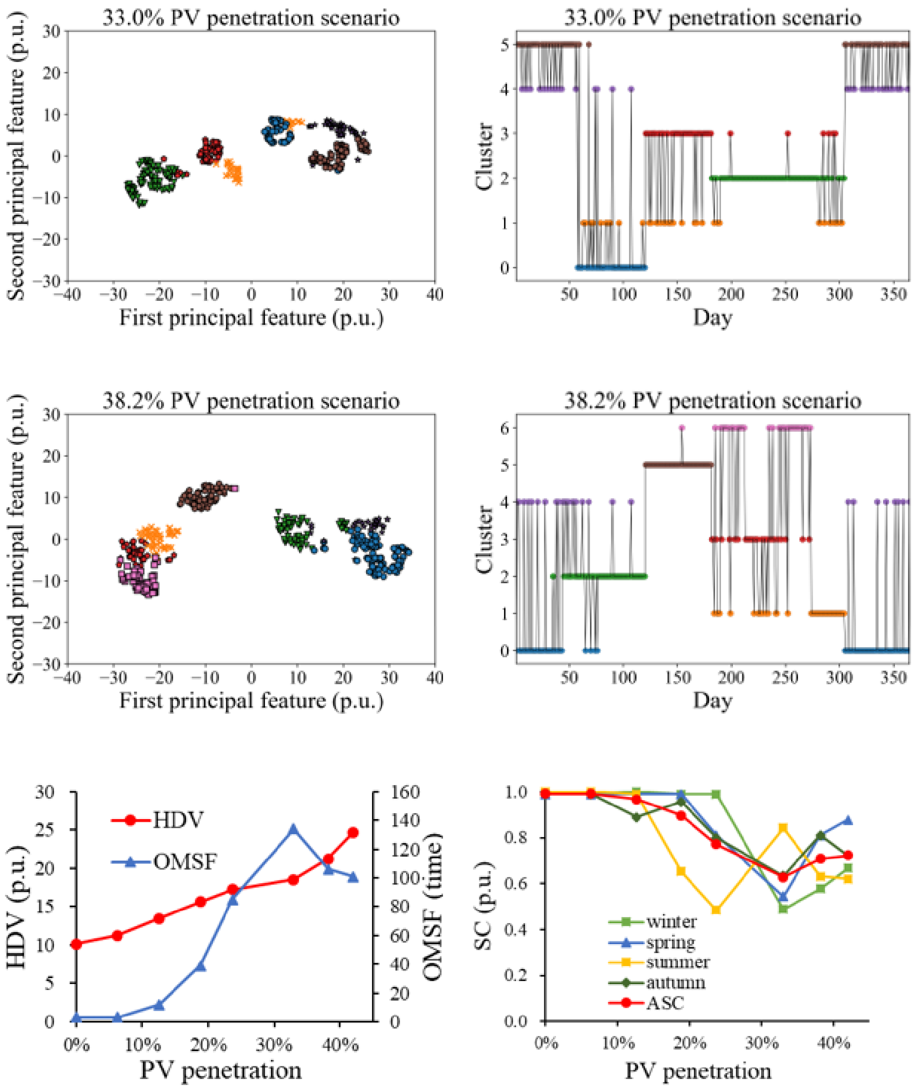


More results

Adding more wind power



Adding more PV power



Conclusions

- Data-driven approach provides more intuitive insight on the diversity of power system operation mode.
- Under low renewable energy penetration, the power system operation mode is dominated by load/hydropower and basically consistent with the season.
- With the growing renewable energy penetration, the power system mode is gradually dominated by intermittent PV and wind power, indicating more representative modes are necessary for power system planning.
- The break point is system-dependent, normally when the VRE penetration is higher than 20%~30%.
- The impact of wind power and PV is distinct. Less daily difference are observed when PV penetration is higher than 30%.

References

- [1] Liu Y, Zhang N, Wang Y, et al. Data-Driven Power Flow Linearization: A Regression Approach[J]. IEEE Trans. Smart Grid, 2018.
- [2] Hou Q, Du E, Zhang N, Kang C. Impact of High Renewable Penetration on the Power System Operation Mode: A Data-Driven Approach[J]. IEEE Trans. Power Systems, 2020.
- [3] 侯庆春, 杜尔顺, 田旭, 刘飞, 张宁, 康重庆. 数据驱动的电力系统运行方式分析[J].电机工程学报, 2020.
- [4] Hou Q, Zhang N, Du E , et al. Probabilistic duck curve in high PV penetration power system: Concept, modeling, and empirical analysis in China[J].Applied Energy, 2019.

Q&A