

# Data Classification and Parameter Identification in Power Systems by Manifold Learning

Andrija T. Sarić

Dept. of Power, Electronics & Communication Engineering  
Faculty of Technical Sciences  
Novi Sad, Serbia  
asaric@uns.ac.rs

Mark K. Transtrum

Dept. of Physics & Astronomy  
Brigham Young University  
Provo, UT, USA  
mktranstrum@byu.edu

Aleksandar M. Stanković

Dept. of Electrical and Computer Engineering  
Tufts University  
Medford, MA, USA  
astankov@ece.tufts.edu

**Abstract**—This paper describes a manifold learning algorithm for big data classification and parameter identification in real-time operation of power systems. We assume a black-box setting, where only SCADA-based measurements at the point of interest are available. Data classification is based on diffusion maps, where an improved data-informed metric construction for partition trees is used. Data reduction is demonstrated on an hourly measurement tensor example, collected from the power flow solutions calculated for daily load/generation profiles. Parameter identification is performed on the same example, generated via randomly selected input parameters. The proposed method is illustrated on the case of the static part (ZIP) of a detailed WECC load model, connected to a single bus of a real-world 441-bus power system.

**Index Terms**-- Big data classification, Parameter identification, Manifold learning, Diffusion maps, WECC load model.

## I. INTRODUCTION

Efficient processing of massive high-dimensional data sets is a contemporary challenge in power systems analytics. Many classical data processing algorithms have computational complexity that scales exponentially with the number of dimensions ("curse of dimensionality") [1]. Researchers have proposed different methods for quick extraction of useful information from large datasets to improve the reliability, efficiency, and flexibility of the grid [2]-[5]. However, most variables that generate data points are typically correlated (locally or globally), endowing such data set with low intrinsic dimensionality. It makes sense then to search for a low-dimensional representation of observation (measurement) samples. Correlations between variables might only be local, in which case classical global dimension reduction methods (such as Principal Component Analysis and Multidimensional Scaling) are not suitable for efficient dimension reduction.

Diffusion maps applied in the context of manifold learning are increasingly used to overcome such problems [6]. These nonlinear techniques have the potential to significantly reduce the data dimensionality. However, they are sensitive to the way the data points were sampled. More precisely, if the data are assumed to approximately lie on a manifold, then an eigenmap representation (which builds a graph from neighborhood information) depends on the density of the points on the manifold [6]. This is important when data are produced by the same source but acquired with different sensors and need to be merged or when different sampling rates are present. Both situations are typical in power systems. In these cases, it is necessary to have a canonical representation of the data that retains the intrinsic constraints of the samples (for example, provided by a manifold geometry). Another important issue is data matching to establish a correspondence between two sets obtained from the same source. This is performed by creating partition trees along different axes.

In this paper, data-driven partition trees are used, applying multilevel partitions along two axes (variable/parameter and time) with a basis function defined for each folder [7]. The main assumption in manifold learning and diffusion maps-based algorithms is that data are constrained to lie on or around a low-dimensional manifold [7]. Reference [8] presents an application of diffusion maps to electromagnetic transient analysis in power systems.

This paper focuses on a black-box setting in which only SCADA-based measurements are available. These are obtained from asynchronous remote terminal units (RTUs), typically collected every 2 to 10 seconds. Real-time variations of input parameters are unknown. Data classification and parameter identification (with extension to model reduction) are considered for the static ZIP load example, which is a part of a more complex WECC load model [9]-[11]. It is assumed that only the initial parameters of the ZIP load are known (for

---

This work has been supported by CURENT Engineering Research Center of the National Science Foundation and the Department of Energy under NSF Award Number EEC-1041877, and in part by the Ministry of Education and Science of the Republic of Serbia, under project III-42004.

example, from past load studies), while their real-time variations are unknown. These parameters are estimated in the real-time using available measurements.

The paper is organized as follows: Section II provides a description of the power system model, measurement and parameter tensors, and the nonlinear optimization for identifying ZIP load parameters. Section III classifies data using the manifold learning algorithm. In Section IV, the proposed method is applied to the ZIP part of a detailed WECC load model connected to the single bus in a 441-bus real-world test system in. Finally Section V presents conclusions.

## II. MODEL DESCRIPTION

A data-driven framework for classification of time-dependent measurements will be demonstrated on a power system example, described by nonlinear algebraic equations:

$$\mathbf{0} = \mathbf{g}(\mathbf{z}, \mathbf{p}, t) \quad (1)$$

where  $\mathbf{z}$  is the vector of algebraic variables,  $\mathbf{p}$  is the vector of parameters, and  $t$  is the (scalar) time variable. System measurements are assumed to be of the form

$$\mathbf{y} = \mathbf{h}(\mathbf{z}, \mathbf{p}, t) \quad (2)$$

The evolution ( $\mathbf{g}$ ) and measurement ( $\mathbf{h}$ ) functions, as well as algebraic variables ( $\mathbf{z}$ ) in the power system are unknown in a black-box setting. The only available information are the measurements ( $\mathbf{y}$ ), which are labeled by time ( $t$ ), and the initial values of parameters ( $\mathbf{p}$ ). The goal is to characterize the power system behavior by systematically organizing the observations of its outputs ( $\mathbf{y}$ ).

In this paper it is assumed that at the bus where the WECC model is connected, the SCADA-based measurements are available. This means that active/reactive ( $P/Q$ ) load and voltage magnitude ( $V$ ) measurements are recorded with SCADA-based time stamps. In this local setup the availability of measurements from the rest of the power system is of no interest, since there is no available information about the full power system model (in our black-box environment). Thus, the available measurements are directly used for optimization of the ZIP load model. The time frame for data classification can be, for example, one hour to detect slow variations of daily load/generation profiles.

### A. Measurement and parameter tensors

The measurement ( $\mathbf{Y}$ ) and parameter ( $\mathbf{Z}$ ) tensors are defined as follows. Formally, let  $\mathcal{M}$  denote an ensemble of  $N_m$  sets of observations (measurements) and  $\mathcal{W}$  denote an ensemble of  $N_w$  sets of time windows for observations. Also, let  $\mathcal{P}$  denote an ensemble of  $N_p$  sets of ZIP-based load parameters ( $\mathbf{p}$ )—see the Appendix, where  $\mathbf{p} = (\mathbf{p}_p, \mathbf{p}_q)$  and  $\mathbf{p}_p = (p_{1c}, p_{1e}, p_{2c}, p_{2e}, P_{s0})$ ,  $\mathbf{p}_q = (q_{1c}, q_{1e}, q_{2c}, q_{2e}, Q_{s0})$ . The active and reactive load-frequency dependencies are neglected (i.e.,  $p_{frq}$  and  $q_{frq}$  in (A1,2) are zero). For each  $\mathbf{m} \in \mathcal{M}$  and  $\mathbf{w} \in \mathcal{W}$  there is an observed trajectory  $Y(\mathbf{m}, \mathbf{w}, t)$  of length  $N_t$  of the system variable, where  $t = 1, 2, \dots, N_t$  denotes the time samples.

Let  $\mathbf{Y}$  denote the entire 3D tensor of observations (measurements)  $Y(\mathbf{m}, \mathbf{w}, t)$ , with dimensions  $N_m \times N_w \times N_t$ , where  $N_m$  denotes the number of measurements and  $N_w$

denotes the number of observation time windows (in our case the hourly load variations). Also, let  $\mathbf{Z}$  denote the entire 3D tensor of ZIP-based load parameters,  $Z(\mathbf{m}, \mathbf{w}, t)$ ,  $\mathbf{p} \in \mathcal{P}$  with dimensions  $N_p \times N_w \times N_t$ , where  $N_p$  denotes the number of parameters (in this case, at most 5 + 5 when considering  $\mathbf{p}_p$  and  $\mathbf{p}_q$  simultaneously).

### B. Parameter identification for the ZIP-based load model

Parameters in the ZIP-based active load model are fit to the observation data based on the optimization for time windows  $\mathbf{w} = 1, 2, \dots, N_w$  and time stamps  $t = 1, 2, \dots, N_t$

$$\hat{\mathbf{p}}_{p,wt} = \min \left\{ \|\mathbf{p}_{p,wt} - \mathbf{p}_p\|_1 \right\} \quad (3)$$

subject to

$$\hat{\mathbf{p}}_{p,wt} = P_{s0,wt} \left[ p_{1c,wt} \left( \frac{V_{wt}}{V_0} \right)^{p_{1e,wt}} + p_{2c,wt} \left( \frac{V_{wt}}{V_0} \right)^{p_{2e,wt}} + p_{3c,wt} \right] \quad (4)$$

where

$$P_{s0,wt} = P_{lf,wt} (1 - F_{mA} - F_{mB} - F_{mC} - F_{mD} - F_{el}) \quad (5)$$

and the reference parameter values (elements of  $\mathbf{p}_p$ ) are given in the Appendix. A similar optimization is used for fitting parameters of the ZIP reactive load model.

## III. DATA CLASSIFICATION BY THE MANIFOLD LEARNING ALGORITHM

For each of the  $N_w$  vectors in  $\mathbf{w} \in \mathcal{W}$ , define a trajectory [12]

$$\mathbf{y}_w = \{Y(\mathbf{m}, \mathbf{w}, t) | \forall \mathbf{m}, \forall t\}, \mathbf{w} \in \mathcal{W} \quad (6)$$

Similarly, define  $\mathbf{y}_m$  and  $\mathbf{y}_t$  to be the samples from the standpoints of the measurement and time axes, respectively

$$\mathbf{y}_m = \{Y(\mathbf{m}, \mathbf{w}, t) | \forall \mathbf{m}, \forall t\}, \mathbf{m} \in \mathcal{M} \quad (7)$$

$$\mathbf{y}_t = \{Y(\mathbf{m}, \mathbf{w}, t) | \forall \mathbf{m}, \forall w\}, t = 1, 2, \dots, N_t, \text{ or } (t \in \mathcal{T}) \quad (8)$$

All data are processed thrice, once for each tensor axis, using an improved, data-informed metric constructed for data-driven partition trees [7], [13].

## IV. APPLICATION

The proposed algorithm for big data classification and parameter identification was tested on a real-world test system—Electric Power Industry of Serbia (a part of the ENTSO-E interconnection), with 441 buses, 280 load buses, 655 branches (transmission lines and two/three winding transformers) and 78 production units.

In-field SCADA-based measurements are unavailable and replaced by solutions to the power flow equations for four different types of daily generation profiles (thermal, hydro, wind, and solar units) and load buses. Part of the generation units (total 14) participate in the Automatic Generation Control (AGC), depending on their participation factors, to compensate for an imbalance between the total load and generation obtained from daily profiles.

Power flow calculations are performed by PSS/E (ver. 33.5.2), in which the WECC load is modeled by the user-defined CMLDBLU1 model. Note that the complete WECC load model is used to calculate the initial conditions of the state

variables, while only the static part of the WECC model is analyzed in more details (see the red rectangle in Figure A1).

#### A. Basic results

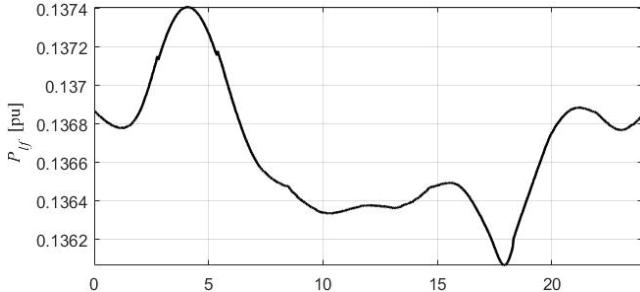
The three-dimensional sets of measurements ( $\mathcal{M}$ ) are generated at the WECC connection point 110 kV (bus 34390, JBOGAT5): active power ( $P_{lf}$ ), reactive power ( $Q_{lf}$ ), and bus voltage ( $V_{lf}$ ). From these values,  $P_{load}$ ,  $Q_{load}$ , and  $V_{load}$  at the load bus are further calculated. Participation of different elements in the WECC load model (A3,4) is assumed to be constant. See Figure A1 for clarification. For daily generation/load profiles, a 24-dimensional set of hourly time windows ( $\mathcal{W}$ ) are generated. For each  $\mathbf{m} \in \mathcal{M}$  and  $\mathbf{w} \in \mathcal{W}$ , observations are made of hourly trajectories of the WECC load model outputs at the connection point for  $N_t = 30$  time stamps per hour (sampled in every 120 sec). A case with more samples will be investigated later in Section IV.B. All of the trajectories are collected into a single 3D measurement tensor ( $\mathbf{Y}$ ), where  $\mathbf{Y} \in \mathbb{R}^{3 \times 24 \times 30}$ , with the system variable trajectories  $Y(\mathbf{m}, \mathbf{w}, t)$  of length  $N_t$  where  $t = 1, 2, \dots, N_t = 30$  denotes the time samples. The daily measurement profile (tensor  $\mathbf{Y}$ ) and extracted hourly profiles are shown in Figure 1.

Similarly, the ten-dimensional set of optimal WECC parameters ( $\mathbf{p} \in \mathcal{P}$ ) are generated. See section II.A for details. For each  $\mathbf{w} \in \mathcal{W}$ , the trajectories of the optimal WECC parameters for  $N_t = 30$  time stamps per hour are observed. All of the trajectories are collected into a single 3D measurement tensor ( $\mathbf{Z}$ ), where  $\mathbf{Z} \in \mathbb{R}^{10 \times 24 \times 30}$ . The ZIP load parameter identification of Eqs. (3)-(4) is performed by adding constraint (4) as a penalizing term to (3) and optimized using the Matlab function ‘fminsearch’. Figure 2 shows daily variations of the optimal ZIP load parameters (due to space limitations, plots for  $p_{1e}$ ,  $q_{1e}$ ,  $p_{2c}$ ,  $q_{2c}$ ,  $p_{2e}$ , and  $q_{2e}$  are omitted—their behavior is similar to that of  $p_{1c}$  and  $q_{1c}$ ). Apart from  $P_{so}$  and  $Q_{so}$ , the variation in all the parameters is small and similar variation for corresponding elements of the tensor is observed.  $P_{so}$  and  $Q_{so}$  follow their corresponding load profiles.

Pattern classification by manifold learning, which involves diffusion geometry with data-driven partition trees, is based on the methodology proposed in [7]. Partition trees and the dominant elements of the eigenvectors along different tensor axes (basic case) are shown in Figures 3 and 4, respectively. From the plot for ‘Time windows ( $\mathbf{w}$ )’ in Figure 4 one can conclude that the hourly patterns in the daily diagram are very different.

#### B. Large-scale tensors

Decreasing the refreshing time step of the SCADA to 10 s, increases  $N_t$  to 360 time stamps per hour and gives a 3D



measurement tensor  $\mathbf{Y} \in \mathbb{R}^{3 \times 24 \times 360}$ . The dominant elements of the eigenvectors for each tensor axis are shown in Figure 5. Comparing Figures 3 and 5 one sees that the number of time stamps negligibly influences the shape of these plots. This suggests that for normal operating conditions (i.e., no major disturbance and generation and loads are subject to small, smooth daily variations) the number of saved time snapshots in the database can be reduced significantly.

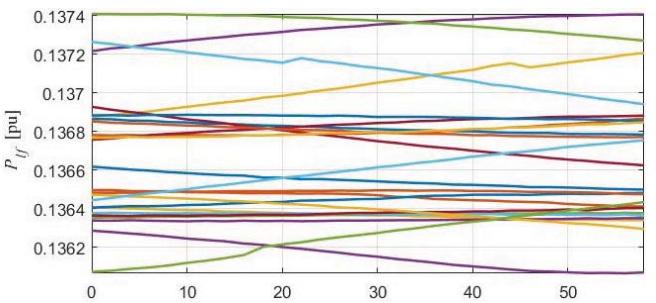
#### C. Parameter identification

For parameter identification (with possible extension to the parameter reduction), the measurement tensor is defined for one time window (for example, the first hour from daily generation/load profiles) as  $\mathbf{Y}'(\mathbf{m}, \mathbf{p}', t)$ , where only the active part of the ZIP load is analyzed. In this case,  $\mathbf{p}' = (\mathbf{p}'_p)$ ,  $\mathbf{p}'_p = (p_{1c}, p_{1e}, p_{2c}, p_{2e})$ ,  $\mathbf{p}' \in \mathcal{P}'$ , where  $\mathcal{P}'$  is the set of uncertain parameter values. Note that  $P_{so}$  is omitted from the set of uncertain parameters because it is directly linked to the measurement set by (A3) and in light of the conclusions from Figure 2.

For the case of SCADA time sampled every 120 s ( $N_t = 30$ ), a set  $\mathcal{P}'$  with  $N_p = 100$  randomly selected parameter values is generated with parameter values drawn from a range  $\pm 20\%$  around the values in the Appendix. All the trajectories are collected into a single 3D measurement tensor ( $\mathbf{Y}'$ ), where  $\mathbf{Y}' \in \mathbb{R}^{3 \times 100 \times 30}$ .

The dominant elements of the eigenvectors for different tensor axes are shown in Figure 6. Plots of parameters for random trials ( $\mathbf{p}'$ ) and times ( $t$ ) show that there are two branches with clear extrema. Also, the variation of the dominant elements of the eigenvectors is very large (in both directions). One can conclude that a relatively small number of random patterns is sufficient to identify the global behavior of eigenvector plots.

An information geometry approach to model reduction of dynamic loads has shown that for a static (ZIP) load model, several parameters are sloppy and do not affect the model behavior [14]. Therefore, it is natural to hold these sloppy parameters fixed. Thus, respecting (A5, 6), three different scenarios of the reduced parameter vector for the active ZIP load can be assumed, as shown in Figure 7. Note that in this figure only plots for the ‘Parameter random trials ( $\mathbf{p}' \in \mathcal{P}'$ )’ axis are shown (with the same axes ranges as in Figure 6), since plots for the ‘Measurements ( $\mathbf{m} \in \mathcal{M}$ )’ and ‘Times ( $t \in \mathcal{T}$ )’ axes are similar to those shown in Figure 6.



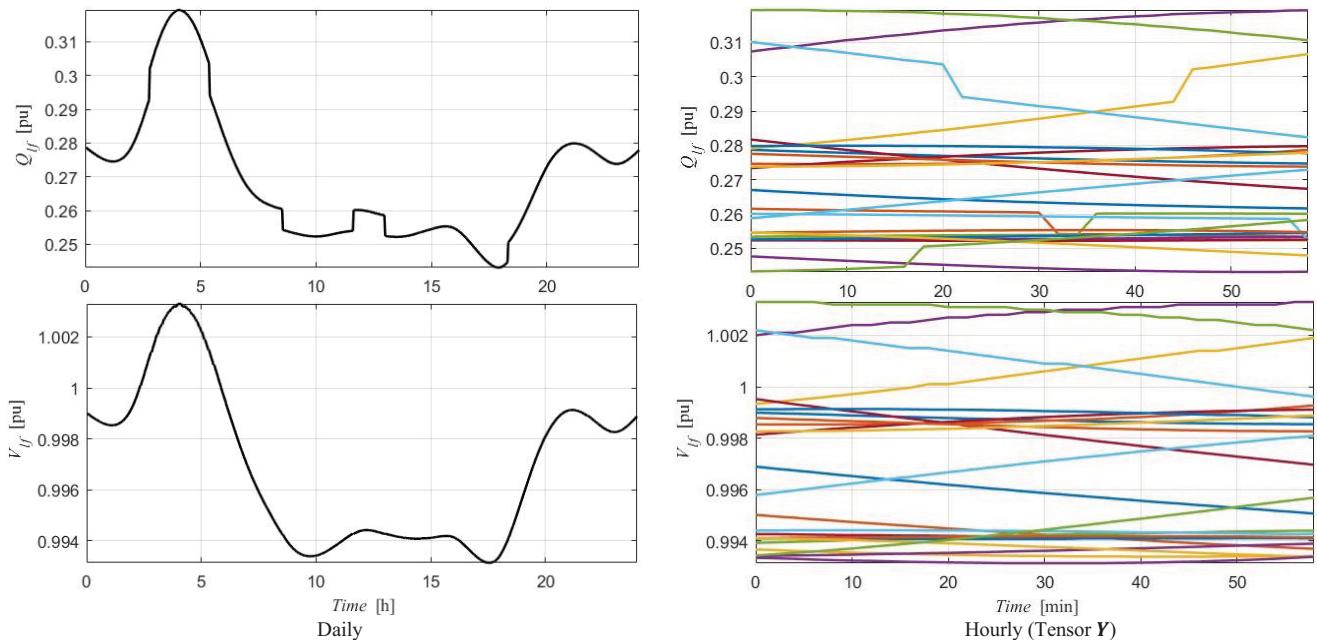


Figure 1. Daily and hourly (tensor  $Y$ ) measurement profiles

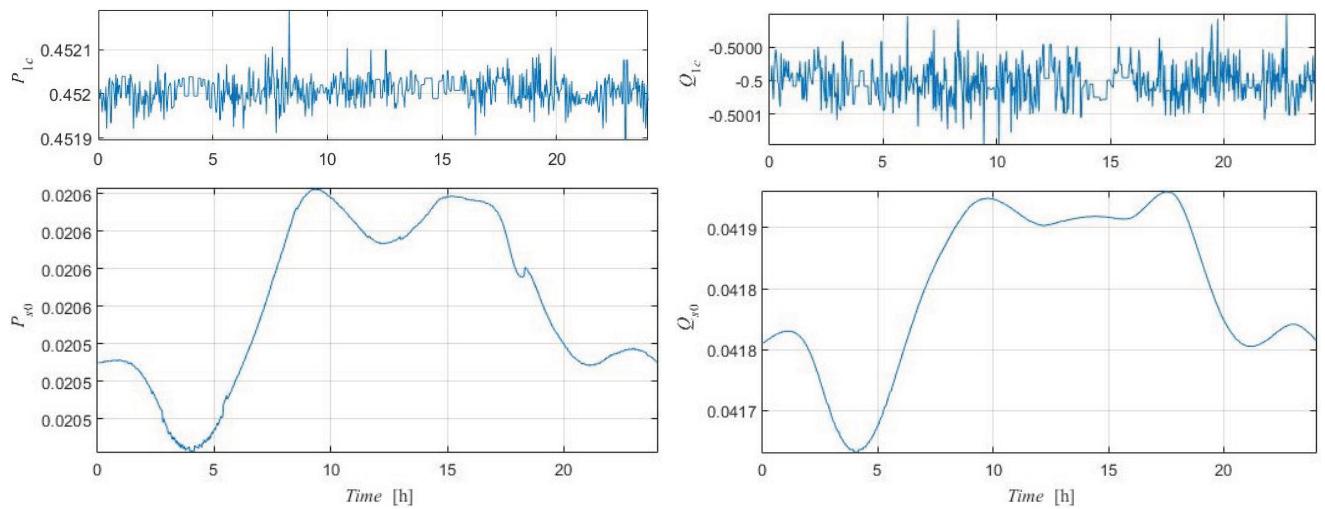


Figure 2. Optimized daily parameters' values

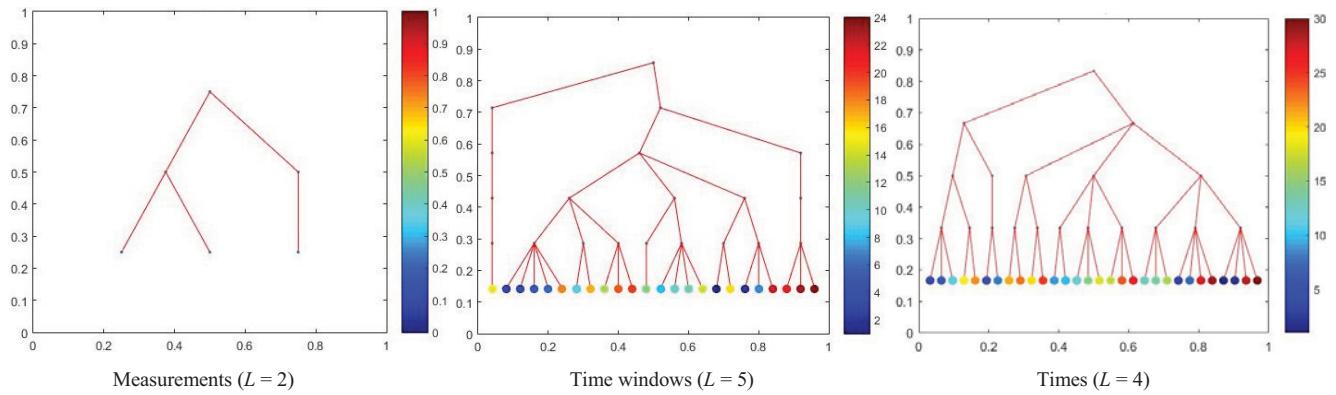


Figure 3. Partition trees in different tensor axes

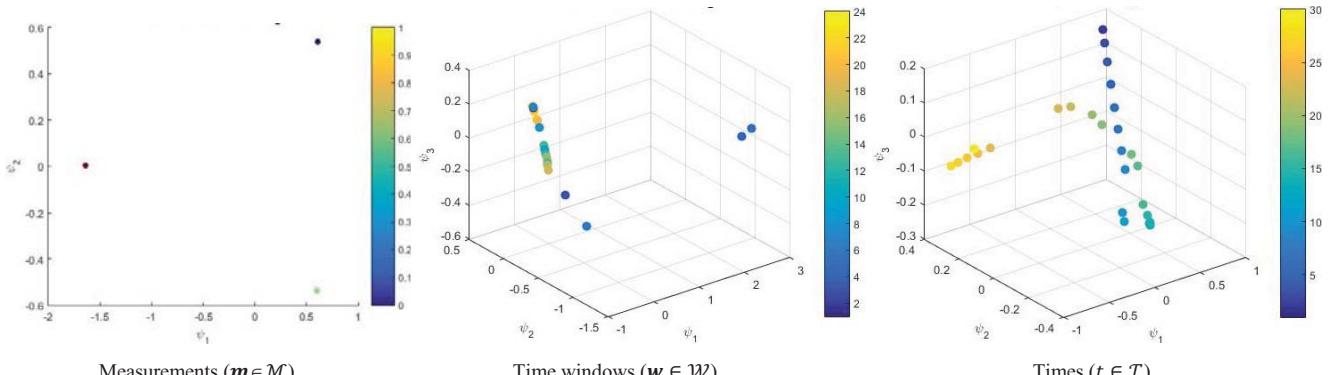


Figure 4. Dominant elements of eigenvectors (basic case) in different tensor's axes

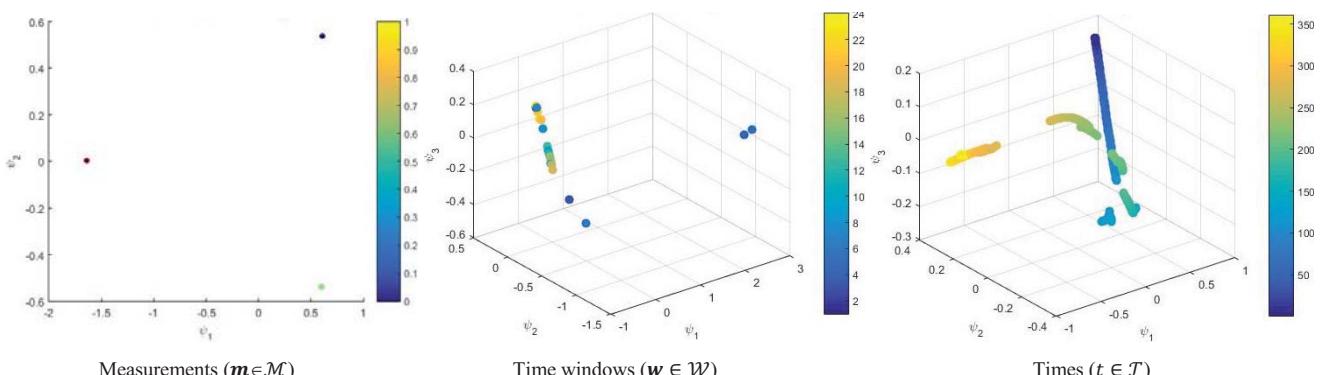


Figure 5. Dominant elements of eigenvectors (large-scale case) in different tensor axes

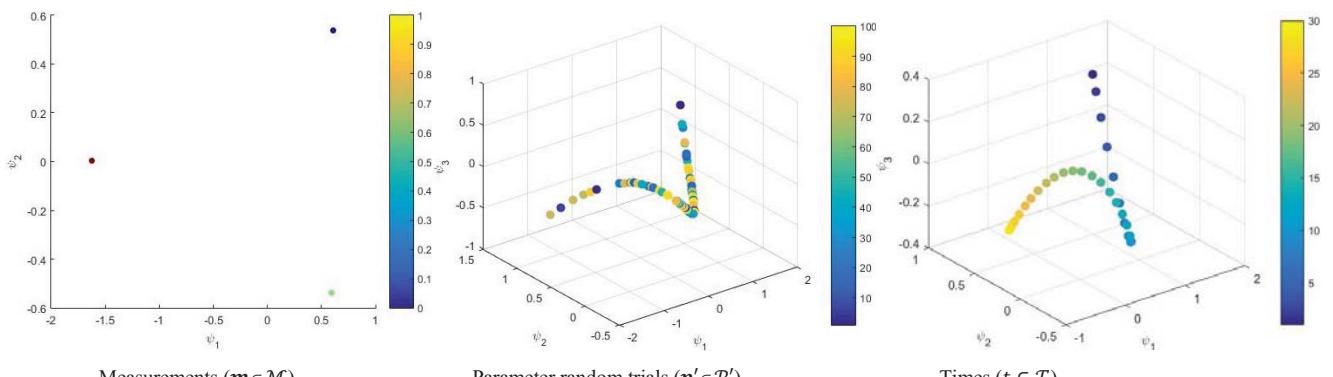


Figure 6. Dominant elements of eigenvectors in different tensor's axes for parameter identification analysis (eight uncertain parameters)

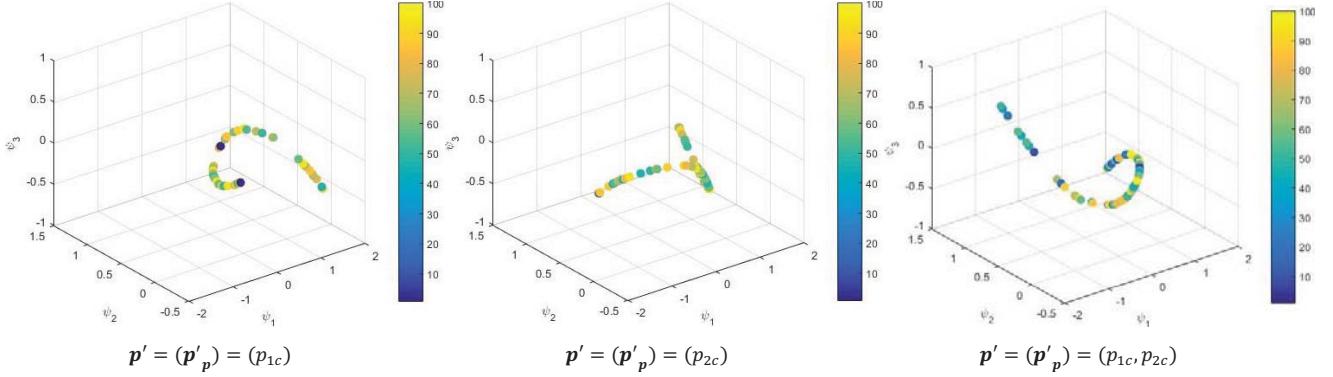


Figure 7. Dominant elements of eigenvectors in parameter random trials ( $\mathbf{p}'$ ) axis for different configurations of uncertain parameters

## V. CONCLUSION

Development of mathematical techniques that operate directly on observations or measurements (i.e., data-driven methods) is of increasing relevance for power systems. One class of such methods by-passes the need to precisely select variables and parameters as well as the need to derive accurate, closed-form equations (i.e., white or gray-box approaches). Initially unorganized measurement data are classified along the dimensions of measurements (inputs), state variables (time windows in our case), parameter settings (inputs) and time snapshot values. Next, a data-informed metric for each type of variation is iteratively constructed. While construction of the partition trees considers the coupling among different tensor axes, the final representation for each entity is separate.

A natural question for this methodology is related to detecting changes in the embedding dimensions with the increased temporal sampling rate. Such dimension changes are important for successful model classification and reduction. The results presented here show that the embedding plot does not change significantly with increased number of time stamps.

Extensions to this study could include: 1) gray-box modeling for SCADA-based measurements, 2) studying the influence of PMUs on data classification and parameter identification (black- and gray-box approaches), and 3) exploring how to enable time-predictive capabilities based on the agnostically organized measurement database (i.e., requiring no knowledge of the mathematical model).

## REFERENCES

- [1] S. Lafon, Y. Keller, and R. R. Coifman, "Fusion and multicue matching by diffusion maps," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1784-1797, Nov. 2006.
- [2] Y. Weng, A. Kumar, M. B. Saleem, and B. Zhang, "Big data and deep learning platform for terabyte-scale renewable datasets," *Proc. of the Power Systems Computation Conference (PSCC)*, Dublin, Ireland, Jun. 11-15, 2018.
- [3] Y. Lokhov, M. Vuffray, D. Shemetov, D. Deka, and M. Chertkov, "Online learning of power transmission dynamics," *Proc. of the Power Systems Computation Conference (PSCC)*, Dublin, Ireland, Jun. 11-15, 2018.
- [4] D. Cao et al., "Design and application of big data platform architecture for typical scenarios of power system," *Proc. of the IEEE PES General Meeting*, Portland, OR, Aug. 5-9, 2018.
- [5] Y. Yu, T. Y. Ji, M. S. Li, and Q. H. Wu, "Short-term load forecasting using deep belief network with empirical mode decomposition and local predictor," *Proc. of the IEEE PES General Meeting*, Portland, OR, Aug. 5-9, 2018.
- [6] R. R. Coifman and S. Lafon, "Diffusion maps", *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5-30, Jul. 2006.
- [7] O. Yair, R. Talmon, R. R. Coifman, and I. G. Kevrekidis, "Reconstruction of normal forms by learning informed observation geometries from data," *Proc. of the National Academy of Sciences of the United States of America (PNAS)*, vol. 114, no. 38, pp. E7865-E7874, Aug. 2017.
- [8] M. C. Arvizu and A. R. Messina, "Dimensionality reduction in transient simulations: A diffusion maps approach," *IEEE Trans. Power Systems*, vol. 31, no. 5, pp. 2379-2389, Oct. 2016.
- [9] D. Kosterev et al., "Load modeling in power system studies: WECC progress update," *Proc. of the IEEE PES General Meeting*, Pittsburgh, PA, Jul. 20-24, 2008.
- [10] WECC Composite Load Model Specifications 01-27-2015. [Online] Available: [https://www.wecc.biz/Reliability/WECC Composite Load Model Specifications 01-27-2015.docx](https://www.wecc.biz/Reliability/WECC%20Composite%20Load%20Model%20Specifications%2001-27-2015.docx)
- [11] G. Thoman, J. Senthil, "A simple CMLD initialization (and comparison of PSSE model initialization with PSLF results)," available: [https://www.wecc.biz/Administrative/InitializeExample\\_051516.docx](https://www.wecc.biz/Administrative/InitializeExample_051516.docx)
- [12] D. W. Sroczynski, O. Yair, R. Talmon, and I. G. Kevrekidis, "Data-driven evolution equation reconstruction for parameter-dependent nonlinear dynamical systems," *Israel Journal of Chemistry*, vol. 58, pp. 1-9, Apr. 2018.
- [13] W. E. Leeb, "Topics in Metric Approximation," *Ph.D. dissertation*, Yale University, New Haven, CT, USA, 2015.
- [14] C. F. Youn, A. T. Sarić, M. K. Transtrum, and A. M. Stanković, "Information geometry for model reduction of dynamic loads in power systems," *Proc. of the IEEE PowerTECH Conference, Session OS25: "Wide Area Monitoring, Protection and Control – WAMPAC II"*, Manchester, United Kingdom, Jun. 18-22, 2017.

## APPENDIX

The static load is part of WECC load model (Figure A1) is represented by algebraic equations as follows:

$$P_s = P_{s0} \left[ p_{1c} \left( \frac{V}{V_0} \right)^{p_{1e}} + p_{2c} \left( \frac{V}{V_0} \right)^{p_{2e}} + p_{3c} \right] (1 + p_{frq} \Delta f) \quad (A1)$$

$$Q_s = Q_{s0} \left[ q_{1c} \left( \frac{V}{V_0} \right)^{q_{1e}} + q_{2c} \left( \frac{V}{V_0} \right)^{q_{2e}} + q_{3c} \right] (1 + q_{frq} \Delta f) \quad (A2)$$

where:

$$P_{s0} = P_{lf} (1 - F_{mA} - F_{mB} - F_{mC} - F_{mD} - F_{el}) \quad (A3)$$

$$Q_{s0} = Q_{lf} (1 - F_{mA} - F_{mB} - F_{mC} - F_{mD} - F_{el}) \quad (A4)$$

$$p_{3c} = 1 - p_{1c} - p_{2c}; q_{3c} = 1 - q_{1c} - q_{2c} \quad (A5)$$

$\Delta f$  – frequency deviation

where input data are as follows:

### Basic power

$$S_{base} = 50 \text{ MVA} ;$$

**Load participations** (Motors A, B, C and D, as well as electronic load, respectively)

$$F_{mA} = 0.2; F_{mB} = 0.2; F_{mC} = 0.1; F_{mD} = 0.25; F_{el} = 0.1.$$

**Static load** (frequency dependence neglected, or  $p_{frq} = 0$  and  $q_{frq} = 0$ )

$$p_{1c} = 0.452; p_{1e} = 2; p_{2c} = 0.548; p_{2e} = 1.$$

$$q_{1c} = -0.5; q_{1e} = 2; q_{2c} = 1.5; q_{2e} = 1.$$

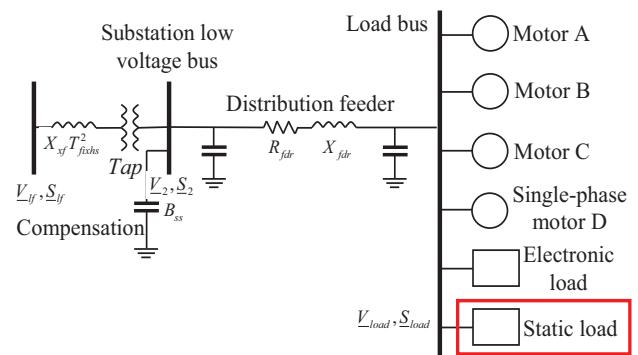


Figure A1. WECC load model