



Dynamic data-driven and model-based recursive analysis for estimation of battery state-of-charge [☆]



Yue Li, Pritthi Chattopadhyay, Sihan Xiong, Asok Ray ^{*}, Christopher D. Rahn

Pennsylvania State University, University Park, PA 16802, USA

HIGHLIGHTS

- A combination of symbolic time series analysis and linear least-squares filtering.
- Markov machine representation for dynamic data-driven analysis.
- Validation on experimental data of pairs of current and voltage data.

ARTICLE INFO

Article history:

Received 7 April 2016

Received in revised form 26 September 2016

Accepted 8 October 2016

Keywords:

Battery state of charge

Dynamic data-driven application systems

Symbolic time series analysis

Recursive Bayesian filtering

ABSTRACT

This paper addresses estimation of battery state-of-charge (SOC) from the joint perspectives of dynamic data-driven and model-based recursive analysis. The proposed SOC estimation algorithm is built upon the concepts of symbolic time series analysis (STSA) and recursive Bayesian filtering (RBF) that is a generalization of the conventional Kalman filtering. A special class of Markov models, called $\times D$ -Markov (pronounced as *cross D*-Markov) machine, is constructed from a symbolized time-series pair of input current and output voltage. A measurement model of SOC is developed based on the features obtained from the $\times D$ -Markov machine. Then, a combination of this measurement model and a low-order model of the SOC process dynamics is used for construction of the RBF. The proposed algorithm of SOC estimation has been validated on (approximately periodic) experimental data of (synchronized) current-voltage time series, generated from a commercial-scale lead-acid battery system.

© 2016 Published by Elsevier Ltd.

1. Introduction

The state-of-charge (SOC) indicates the maximum charge that can be drawn from a battery at a given instant of time; SOC is a critical system state for battery operation from both safety and performance perspectives. Since the state of the art in battery technology does not permit direct measurements of SOC, the challenge is to perform accurate and reliable estimation of SOC in real time based on the time series of input current and output voltage data. Many model-based filtering techniques have been reported on the topic of SOC estimation [1,2]. In general, state-space (e.g., equivalent circuit [3] and electrochemical [4]) models are constructed based on the understanding of the physics and chemistry of battery

cells. In these models, the input is the charging/discharging current, the output is the battery voltage, and SOC is a model state.

To achieve a desired level of estimation accuracy in the presence of nonlinearities of battery dynamics, several algorithms have been proposed based on the concept of Kalman filtering; examples are: extended Kalman filtering (EKF) [5], adaptive extended Kalman filtering (AEKF) [6–8], and unscented Kalman filtering (UKF) [9,10]. [11] uses a dual particle filter to simultaneously predict SOC and drift current estimation. However, such techniques may not result in accurate estimation of the battery state due to the lack of adequate understanding of electrochemical dynamics of batteries. This shortcoming of model-based filtering could be complemented by dynamic data-driven application systems (DDDAS) [12] that make use of the information derived from the measurement ensemble, as a result of which they do not require detailed knowledge of battery electro-chemistry; therefore DDDAS could be more efficient in dealing with computational complexity (e.g., execution time and memory requirements) [13,14].

With the recent development of sensing techniques and sophisticated testing platforms, large volumes of high-quality data could

[☆] This work has been supported in part by the U.S. Air Force Office of Scientific Research (AFOSR) under Grant No. FA9550-15-1-0400. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

^{*} Corresponding author.

E-mail addresses: yol5214@psu.edu (Y. Li), pxc271@psu.edu (P. Chattopadhyay), sux101@psu.edu (S. Xiong), axr2@psu.edu (A. Ray), cdrahn@psu.edu (C.D. Rahn).

Nomenclature

C	battery capacity	\tilde{y}_i	symbolized voltage output at the i th instant
D	depth of the Markov machine	x_k	state of charge at the end of the k th time window
k	k th time window	Ψ_k	depth of discharge in the k th time window
n_k	total instants at the end of k th window	Π	probability morph matrix of a probabilistic finite state automata
\mathcal{Q}	set of states of a probabilistic finite state automata	Σ	alphabet set of a probabilistic finite state automata
$p(x_k x_{k-1}; \Psi_k)$	dynamic model of SOC (probability of current state given past state and input)		
$p(y_k x_k; \mathbb{U}_k)$	measurement model of SOC (probability of current measurement given current state)	<i>Acronym</i>	
\mathbb{U}_k	input symbol string corresponding to the k th time window	DOD	depth of discharge
u_i	real-valued current input at the i th instant	MEP	maximum entropy partitioning
\tilde{u}_i	symbolized current input at the i th instant	PFSA	probabilistic finite state automata
\mathbb{Y}_k	output symbol string corresponding to the k th time window	RBF	recursive bayesian filtering
y_i	real-valued voltage output at the i th instant	SOC	state of charge
		STSA	symbolic time series analysis

be made available to cover wide ranges of operating conditions. This situation leads to the possibility of formulating robust and accurate data-driven SOC estimation algorithms (e.g., k -nearest neighbor (k -NN) regression [15], support vector machines (SVM) [16], and artificial neural networks (ANNs) [17]) that can be applied to battery management systems in practice. However, there exists common problems for purely data-driven methods, namely, over-fitting and under-fitting of data, which may compromise the performance of the SOC estimation algorithm.

To overcome the individual deficiencies of the model-based and data-driven approaches, several hybrid methods have been proposed by combining both the model-based and data-driven approaches. For example, Charkhgard and Farrokhi [18] used neural networks to approximate a state-space model and combined it with an extended Kalman filter for SOC estimation. Xu et al. [19] reported a similar approach with a stochastic fuzzy neural network (SFNN). He et al. [20] adopted an unscented Kalman filter (UKF) to enhance the accuracy of nonlinear estimation.

Li et al. [21] reported a dynamic data-driven method for SOC estimation via Markov modeling on the battery voltage time series. The underlying concept was built upon the theory of symbolic analysis [22,23] that extracts the dynamic information from symbolized time series. Along this line, Li et al. [14] have extended the method by considering synchronized inputs (i.e., current) and outputs (i.e., voltage) together in the setting of a Markov model to incorporate input-adaptive features. Later, Sarkar et al. [24] investigated the cross-dependency between symbolized input strings and symbolized output strings using the $\times D$ -Markov (pronounced as *cross D*-Markov) machine for SOC estimation. Recently, Chattopadhyay et al. [25] constructed a sequential SOC estimation framework based on symbolic Markov modeling [14].

This paper is a significant extension of the authors' earlier work [24,25] and proposes a framework for recursive estimation of SOC by combining the underlying concepts of $\times D$ -Markov machines and recursive Bayesian filtering (RBF). In this work, the causal cross-dependence between the battery input current and output voltage is modeled and represented as a $\times D$ -Markov machine, which is somewhat analogous to a transfer function in the sense that it analyzes the relation between the input and output of the dynamical system. Then, the SOC measurement model is formulated by a kernel-based conditional density estimation, which is constructed from the $\times D$ -Markov features obtained in the training phase. Finally, a model-based recursive Bayesian estimator is constructed, which filters out the outliers in SOC estimation to reduce the estimation error. The proposed method has been validated

using experimental data of a (commercial-scale) lead-acid battery under varying input conditions.

From the perspectives of state-of-the-art battery technology, the work reported in this paper has the following major contributions:

- Representation of the nonlinear dynamics of a battery by symbolic Markov machine models that have significantly reduced computational complexity relative to the commonly used physical dynamic models.
- Development of an SOC measurement model based on a variant of statistical regression. The probability of SOC being a certain value, conditioned on the battery output and input is estimated as a locally weighted average using kernel as a weighing function.
- Incorporation of the dynamic data-driven measurement model within the structure of a model-based RBF for real-time SOC estimation.
- Validation of the proposed algorithms on experimental data collected from a commercial-scale lead-acid battery.

Although the work reported in this paper focuses on lead-acid batteries, the proposed methodology can be extended to other types of batteries (e.g., Li-ion and Ni-MH).

The paper is organized in five sections including the present one. Section 2 briefly describes the underlying principles of symbolic Markov modeling that is used for development of a SOC measurement model. Section 3 proposes a hybrid (i.e., combined dynamic data-driven and model-based recursive) SOC estimation algorithm by combining the data-driven SOC measurement model and model-based RBF. Section 4 briefly presents the details of the experiment performed for acquisition of battery system data as well as the results of the proposed estimation algorithm on the experimental data. Finally, the paper is summarized and concluded in Section 5 along with recommendations for future research.

2. Symbolic Markov modeling

This section presents a general framework of symbolic $\times D$ -Markov modeling [24] for analytical measurement of battery SOC. First, the time series of (synchronized) input (i.e., current) and output (i.e., voltage) are preprocessed and discretized into symbol strings via maximum entropy partitioning [26]. Then, $\times D$ -Markov machines are constructed to obtain features that

represent the dynamical behavior of the battery at different SOC conditions.

2.1. Time series symbolization

This subsection addresses the procedure for time series symbolization. First, battery input current and output voltage are individually normalized into zero-mean and unit-variance time series as:

$$\tilde{\theta}_i = \frac{\theta_i - \mu_{\theta}}{\sigma_{\theta}} \text{ for } i = 1, 2, \dots, L \quad (1)$$

where μ_{θ} is the mean and σ_{θ} is the standard deviation of the unprocessed time series $\theta_1, \theta_2, \dots, \theta_L$ of finite length L . The objective of pre-processing is to remove the undesirable effects due to bias (e.g., due to measurement drifts) and variance fluctuation (e.g., due to different noise level) from time series before symbolization.

The first step in symbolization of time series involves partitioning of the range space of the signal. The signal space is partitioned into a finite number of cells and a symbol associated with each cell, i.e., the number of cells is identically equal to the cardinality $|\Sigma|$ of the (symbol) alphabet Σ . As an example, for the one-dimensional time series in Fig. 1, three partitioning lines divide the ordinate (i.e., y-axis) of the time series profile into four mutually exclusive and exhaustive regions that form a partition of the range space, where each region is labeled with one symbol from the alphabet Σ . If the value of time series at a given instant is located in a particular cell, then it is coded with the symbol associated with that cell. In this way, a symbol string is generated from the (finite-length) time series. Details are reported in [22].

Ensembles of time series data have been symbolized by using a partitioning tool, called maximum entropy partitioning (MEP) [26], that maximizes the entropy of the generated symbols; therefore, the information-rich cells of a data set are partitioned finer and those with sparse information are partitioned coarser (i.e., each cell contains approximately equal number of data points). The choice of alphabet size $|\Sigma|$ largely depends on the specific data set and the allowable loss of information [23].

2.2. $\times D$ -Markov modelling

While the details are reported in authors' earlier publications [22–24], this subsection very concisely presents the pertinent concepts for construction of D -Markov machine and $\times D$ -Markov machine models.

Definition 2.1 (DFA [27]). A deterministic finite-state automaton (DFA) G is a triple (Σ, Q, δ) where:

- Σ is a (nonempty) finite alphabet with cardinality $|\Sigma|$;
- Q is a (nonempty) finite set of states with cardinality $|Q|$;
- $\delta : Q \times \Sigma \rightarrow Q$ is the state transition map.

Definition 2.2 (Symbol Block). A symbol block, also called a word, is a finite-length string of symbols s_i belonging to the alphabet Σ , where the length of a word $w \triangleq s_1s_2 \dots s_\ell$ with $s_i \in \Sigma$ is $|w| = \ell$, and the length of the empty word ϵ is $|\epsilon| = 0$. The parameters of DFA are extended as:

- The set of all words constructed from symbols in Σ , including the empty word ϵ , is denoted as Σ^* ,
- The set of all words, whose suffix (respectively, prefix) is the word w , is denoted as Σ^*w (respectively, $w\Sigma^*$).
- The set of all words of (finite) length ℓ , where $\ell > 0$, is denoted as Σ^ℓ .

Definition 2.3 (PFSA [22,23]). A probabilistic finite-state automaton (PFSA) is constructed on the algebraic structure of deterministic finite-state automata (DFA) $G = (\Sigma, Q, \delta)$ as a pair $K = (G, \Pi)$, i.e., the PFSA K is a 4-tuple $K = (\Sigma, Q, \delta, \Pi)$, where:

- (1) Σ is a non-empty finite set, called the symbol alphabet, with cardinality $|\Sigma| < \infty$;
- (2) Q is a non-empty finite set, called the set of states, with cardinality $|Q| < \infty$;
- (3) $\delta : Q \times \Sigma \rightarrow Q$ is the state transition map;
- (4) $\Pi : Q \times \Sigma \rightarrow [0, 1]$ is the symbol generation matrix (also called probability morph matrix) that satisfies the condition $\sum_{j=1}^{|\Sigma|} \Pi(i, j) = 1 \quad \forall q_i \in Q$, and $\Pi(i, j)$ is the probability of emission of a symbol $\sigma_j \in \Sigma$ from the state $q_i \in Q$.

Definition 2.4 (D -Markov Machine [22,23]). Let $S = \dots s_{k-1} s_k s_{k+1} \dots$ be a discrete symbol sequences which is a statistically stationary stochastic process. The probability of occurrence of a new symbol depends only on the last D symbols, i.e.,

$$P[s_k | \dots s_{k-D} \dots s_{k-1}] = P[s_k | s_{k-D} \dots s_{k-1}] \quad (2)$$

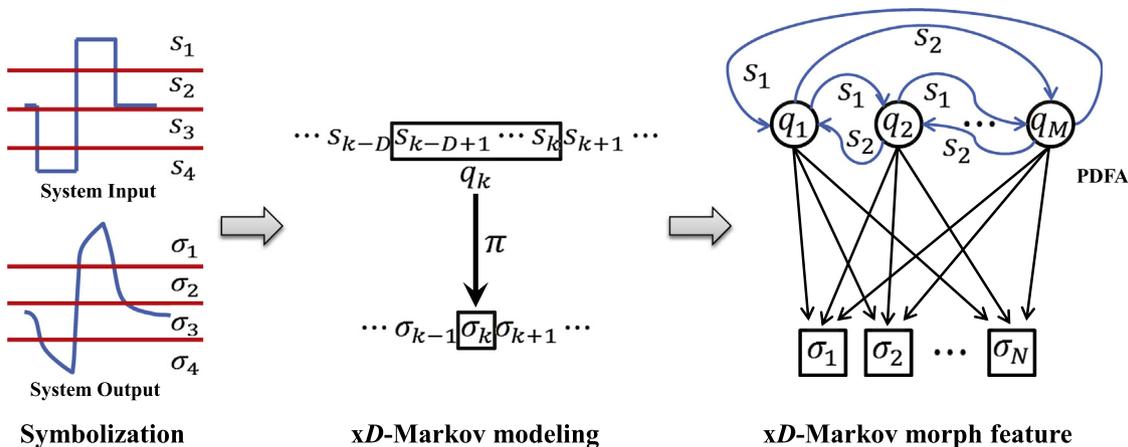


Fig. 1. $\times D$ -Markov (pronounced as cross D -Markov) modeling and feature extraction.

where D is called the depth of the Markov machine. Then, a D -Markov machine is modeled as a PFSA, in which each state is represented by a finite history of D symbols, and is defined as a 4-tuple $\mathcal{M} \triangleq (\mathcal{Q}, \Sigma, \delta, \Pi)$ such that:

- (1) $\Sigma = \{s_1, \dots, s_{|\Sigma|}\}$ is the alphabet of symbols and $|\Sigma| < \infty$ is cardinality of the alphabet.
- (2) $\mathcal{Q} = \{q_1, q_2, \dots, q_{|\mathcal{Q}|}\}$ is the state set with cardinality $|\mathcal{Q}| \leq |\Sigma|^D$, i.e., the states are represented by equivalence classes of symbol blocks of maximum length D corresponding to a symbol sequence \mathbb{S} .
- (3) $\delta: \mathcal{Q} \times \Sigma \rightarrow \mathcal{Q}$ is the state transition mapping, which generates the symbol sequences;
- (4) $\Pi: \mathcal{Q} \times \Sigma \rightarrow [0, 1]$ is the morph matrix of size $|\mathcal{Q}| \times |\Sigma|$; the ij th element $\Pi(i, j)$ of the matrix Π denotes the probability of finding the symbol σ_j at next time step while making a transition from the state q_i .

Thus, the symbol sequence is compressed as a PFSA by assigning the states as words of finite length from the symbol sequence. The PFSA induces a D -Markov chain of finite order [22,23].

Next the notion of $\times D$ -Markov machine is introduced, and the underlying concept is illustrated in Fig. 1.

Definition 2.5 ($\times D$ -Markov Machine [24]). Let $\mathbb{U} \triangleq \dots \tilde{u}_{k-1} \tilde{u}_k \tilde{u}_{k+1} \dots$ and $\mathbb{Y} \triangleq \dots \tilde{y}_{k-1} \tilde{y}_k \tilde{y}_{k+1} \dots$ be two (time-synchronized) stochastic symbol sequences. Then, a $\times D$ -Markov machine, where the Markov assumption holds for \mathbb{Y} with respect to \mathbb{U} , is defined as a 5-tuple $\mathcal{M}_{1 \rightarrow 2} \triangleq (\mathcal{Q}_1, \Sigma_1, \Sigma_2, \delta_1, \Pi_{1 \rightarrow 2})$ such that:

- (1) $\Sigma_1 = \{s_1, \dots, s_{|\Sigma_1|}\}$ is the alphabet of symbols in \mathbb{U} .
- (2) $\Sigma_2 = \{\sigma_1, \dots, \sigma_{|\Sigma_2|}\}$ is the alphabet of symbols in \mathbb{Y} .
- (3) $\mathcal{Q}_1 = \{q_1, q_2, \dots, q_{|\mathcal{Q}_1|}\}$ is the set of states corresponding to \mathbb{U} .
- (4) $\delta_1: \mathcal{Q}_1 \times \Sigma_1 \rightarrow \mathcal{Q}_1$ is the state transition mapping. It is noted that the PFSA structure is built on \mathbb{U} and thus, the transition map δ_1 generates a sequence of symbols that belong to Σ_1 ; however the Markov assumption holds for \mathbb{Y} on the states inferred from \mathbb{U} .
- (5) $\Pi_{1 \rightarrow 2}: \mathcal{Q}_1 \times \Sigma_2 \rightarrow [0, 1]$ is the \times -morph (pronounced as cross morph) matrix of dimension $|\mathcal{Q}_1| \times |\Sigma_2|$; the ij th element $\Pi(i, j)$ of the \times -morph matrix Π denotes the probability of finding the symbol σ_j in the symbol string \mathbb{Y} at next time step while making a transition from the state q_i of the PFSA constructed from the symbol sequence \mathbb{U} .

Similarly, a 5-tuple $\mathcal{M}_{2 \rightarrow 1} \triangleq (\mathcal{Q}_2, \Sigma_2, \Sigma_1, \delta_2, \Pi_{2 \rightarrow 1})$ yields symbols in Σ_1 from states in \mathcal{Q}_2 .

While a D -Markov Machine captures the temporal dynamics of a single sequence, $\times D$ -Markov machine models capture the causal dependency between two sequences.

2.3. Feature extraction by a $\times D$ -Markov machine

Given a (finite-length) symbol string \mathbb{S} over a (finite) alphabet Σ , there exist several construction algorithms to discover the underlying irreducible PFSA model. For both algorithmic simplicity and computational efficiency, $\times D$ -Markov machines [24] have been adopted in this paper for modeling cross-dependence between two PFSAs. The symbolic time series corresponding to the input current \mathbb{X} and output voltage \mathbb{Y} are modeled as PFSAs' K_1 and K_2 respectively (as described in Definition 2.4).

Let $N(\sigma_j, q_k)$ denote the number of times that a symbol σ_k is generated by PFSA K_2 when the state q_j is observed (as a symbol

string) by PFSA K_1 . The maximum a posteriori probability (MAP) estimate of the symbol emission probability of the PFSA K_2 from PFSA K_1 is obtained by frequency counting [23] as:

$$\Pi_{1 \rightarrow 2}(q_k, \sigma_j) \approx \frac{1 + N(\sigma_j, q_k)}{|\Sigma_2| + \sum_{\ell} N(\sigma_{\ell}, q_k)} \quad (3)$$

If no event is generated at a combination of symbol $\sigma_j \in \Sigma_2$ and state $q_k \in \mathcal{Q}_1$, then there should be no preference to any particular symbol and it is logical to have $\Pi_{1 \rightarrow 2}(k, j) = 1/|\Sigma_2|$. The above procedure guarantees that $\times D$ -Markov machines, constructed from two (finite-length) symbol strings, will have an (elementwise) strictly positive \times -morph matrix $\Pi_{1 \rightarrow 2}$ that is constructed by estimating the individual emission probabilities $\Pi_{1 \rightarrow 2}(k, j)$ for $k \in \{1, 2, \dots, |\mathcal{Q}_1|\}$ and $j \in \{1, 2, \dots, |\Sigma_2|\}$. Similar procedures hold for the \times -morph matrix $\Pi_{2 \rightarrow 1}$.

3. Hybrid recursive SOC estimation

This section first introduces definitions of pertinent battery parameters at a given ambient temperature [28], followed by formulation of a recursive Bayesian estimation filter.

Definition 3.1 (Battery capacity). The capacity $C(T)$ of a battery at time T (in the slow time-scale) is its maximum charge (in units of ampere-hours) that can be drawn from its fully charged condition at a rate $C(T)/30$ (in units of amperes).

Definition 3.2 (DOD and SOC). Let a battery be fully charged at time T (in the slow time-scale) and let $I(\tau)$ be the applied current (in units of amperes) at time τ (in the fast time-scale). Then, depth of discharge (DOD) and state of charge (SOC) at time $T + \Delta T$ are respectively defined as:

$$DOD(T + \Delta T) = \frac{1}{C(T)} \int_T^{T+\Delta T} I(\tau) d\tau, \quad \Delta T \geq 0 \quad (4)$$

$$SOC(T + \Delta T) = 1 - DOD(T + \Delta T), \quad \Delta T \geq 0 \quad (5)$$

If the battery capacity C is a known parameter that is assumed to be a constant for the purpose of SOC estimation, then DOD depends only on the input current that is also a known variable.

Now the issue of recursive Bayesian estimation (also known as a recursive Bayes filter (RBF)) is addressed, which is a general probabilistic approach for estimating unobserved states recursively over time using incoming measurements and a dynamic model. It consists of the following two steps:

- (1) Predicting the state from the previously estimated (or initially provided) state by using a dynamic model.
- (2) Updating the state estimation using the current measurement based on the measurement model.

Fig. 2 depicts the construction of SOC dynamic model and data-driven measurement model used in the implementation of a recursive Bayes filter for SOC estimation. In this filter, the unobserved SOC state and the dynamic model input DOD change only in the slow time scale, while the symbolic measurements are extracted from (synchronized) time series pairs of real-valued current input and voltage output in the fast time scale. The following notations are used in the proposed Bayes filter for SOC state estimation:

- u_i : Real-valued current input at the i th instant;
- y_i : Real-valued voltage output at the i th instant;
- \tilde{u}_i : Real-valued voltage output at the i th instant;

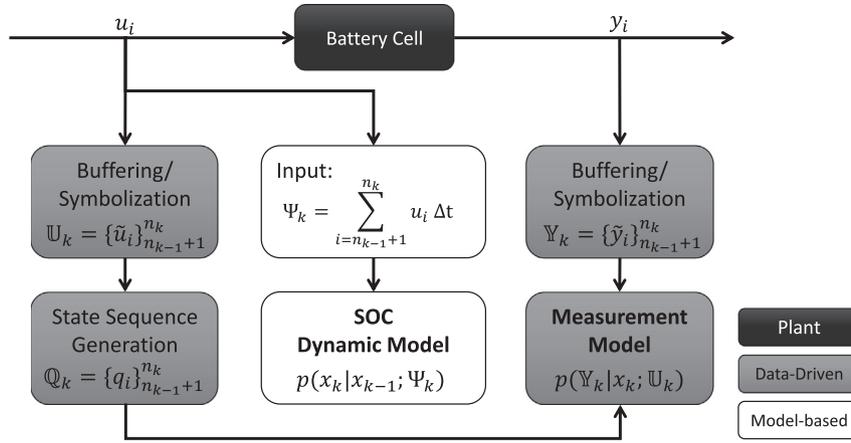


Fig. 2. Schematic representation of SOC dynamic model and measurement model.

- \tilde{y}_i : Symbolized voltage output at the i th instant;
- n_k : The total instants at the end of k th window;
- x_k : State of charge (SOC);
- Ψ_k : Depth of discharge (DOD);
- \mathbb{U}_k : Input symbol string during k th window;
- \mathbb{Y}_k : Input symbol string during k th window;
- $p(x_k|x_{k-1}; \Psi_k)$: SOC Dynamic model;
- $p(\mathbb{Y}_k|x_k; \mathbb{U}_k)$: Data-driven Measurement model;

where $\Psi_k \triangleq \sum_{i=n_{k-1}+1}^{n_k} u_i \Delta t$ and Δt is the sampling time for the fast time scale; $\mathbb{U}_k \triangleq \{\tilde{u}_{n_{k-1}+1}, \dots, \tilde{u}_{n_k}\}$ and $\mathbb{Y}_k \triangleq \{\tilde{y}_{n_{k-1}+1}, \dots, \tilde{y}_{n_k}\}$; and \tilde{u} and \tilde{y} denote the respective input and output symbols corresponding to real-valued inputs u and outputs y in the fast time scale, as depicted in Fig. 2.

3.1. Analytical measurement model

This subsection describes how to obtain the measurement model via statistical regression. The crucial assumption here is that the estimated conditional probability from the training phase preserves its mathematical structure in the testing phase. For example, let there be M training data sets, where each data set consists of a (synchronized) finite-length time series pair of real-valued current input $\{u_k\}_{k=1}^L$ and real-valued voltage output $\{y_k\}_{k=1}^L$ with the associated SOC label x . A $\times D$ -Markov machine is constructed from each data set by the procedure described in Section 2.3. Thus, for each data set i , there exists a symbolic output string $\mathbb{Y}^i = \{\tilde{y}_k^i\}_{k=1}^L$, the PFSA state string \mathbb{Q}^i generated from the symbolic input string $\mathbb{U}^i = \{\tilde{u}_k^i\}_{k=1}^L$, the $\times D$ -Markov morph matrix Π^i , and the SOC value x^i . The cross dependency between the current (input) and voltage (output) is expected to vary for different values of SOC. The proposed algorithm is schematically represented in Fig. 3.

Given the input state string \mathbb{Q} and the cross dependency specified by Π^j (generated from the j th training data), the probability of observing a symbol output string \mathbb{Y} is:

$$p(\mathbb{Y}|\mathbb{U}, \Pi^j) = P(\mathbb{Y}|\mathbb{Q}, \Pi^j) = \prod_{k=1}^{L-D+1} \Pi^j(q_k, \tilde{y}_{k+D-1}) \quad (6)$$

where L is the length of the symbol string and D is the depth for the Markov modeling.

Let \mathbb{U}_k and \mathbb{Y}_k denote the input and output symbol strings, in the k th time window of the testing phase. The probability of observing \mathbb{Y}_k conditioned on the system state x_k and the input \mathbb{U}_k is estimated as:

$$p(\mathbb{Y}_k|x_k; \mathbb{U}_k) = \sum_{i=1}^M \mathcal{W}(x_k, \Pi^i) p(\mathbb{Y}_k|\mathbb{U}_k, \Pi^i) \quad (7)$$

where the weighting function \mathcal{W} is formulated as a (Gaussian) radial basis function kernel:

$$\mathcal{W}(x_k, \Pi^i, \gamma) = \frac{\exp(-\gamma(x_k - x^i)^2)}{\sum_{j=1}^M \exp(-\gamma(x_k - x^j)^2)} \quad (8)$$

where $\gamma > 0$ is a free parameter and x^i is the SOC label associated with the \times -morph matrix Π^i .

Compared to physical measurements, symbolic measurements are more robust to noise due to the increased signal-to-noise-ratio (SNR) as a consequence of discretization [29], and it provides a better state estimation in the sense that the conditional variance of measurements given the state is reduced.

3.2. Recursive Bayesian filtering for SOC estimation

This subsection combines the data-driven (analytical) measurement model with the process model of SOC in the sense of Bayesian filtering. For this purpose, a linear state-space model of SOC variation is formulated as:

$$\text{State equation: } x_k = x_{k-1} - \frac{1}{C_k} \Psi_k + w_k \quad (9)$$

where C_k is the maximum capacity of the battery during the k th time window, and w_k is the process noise which is assumed to be drawn from a zero-mean Gaussian distribution with variance of v_k . The dynamic variable Ψ_k is the depth of discharge (DOD) from the beginning to the end of k th time window, which is obtained as:

$$\Psi_k \triangleq \sum_{i=n_{k-1}+1}^{n_k} u_i \Delta t \quad (10)$$

where Δt is the sampling time of real-valued current input.

Then, the probability of the current true state x_k given the previous state x_{k-1} and the DOD Ψ_k is assumed to follow a Gaussian distribution $\mathcal{N}(x_{k-1} - \frac{\Psi_k}{C_k}, v_k)$:

$$p(x_k|x_{k-1}; \Psi_k) = f\left(x_k|x_{k-1} - \frac{\Psi_k}{C_k}, v_k\right) = \frac{1}{\sqrt{2\pi v_k}} \exp\left(-\frac{\left(x_k - \left(x_{k-1} - \frac{\Psi_k}{C_k}\right)\right)^2}{2v_k}\right) \quad (11)$$

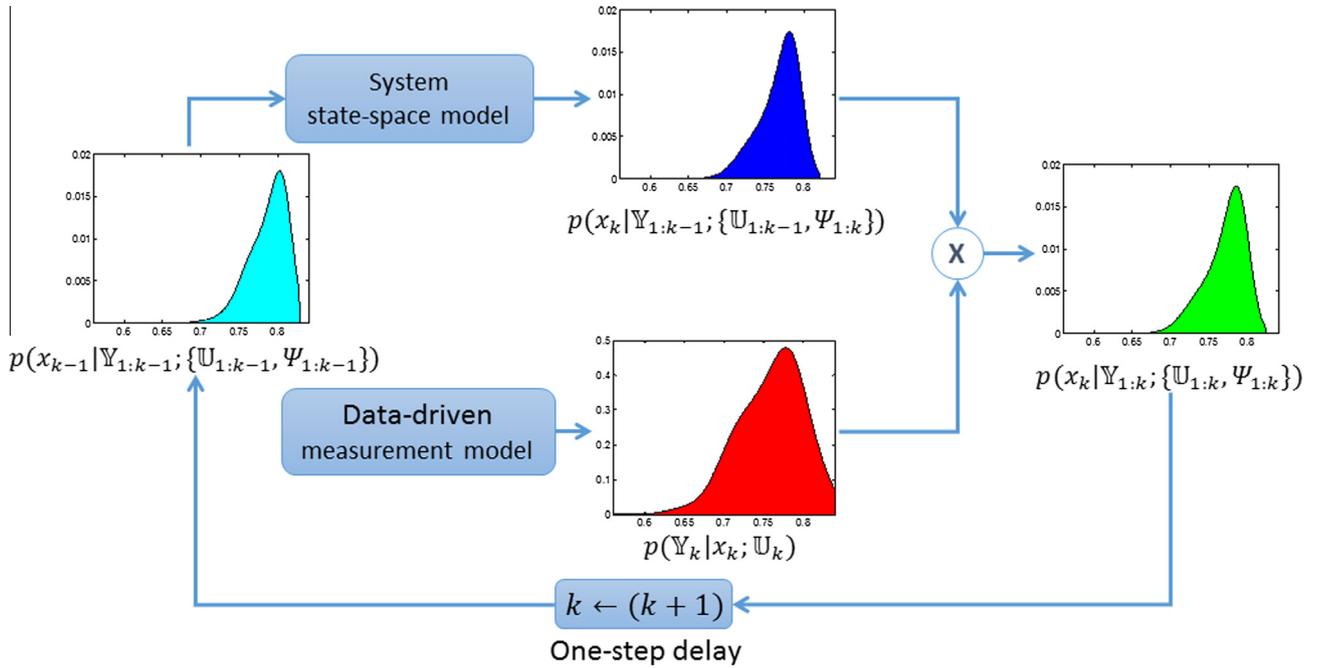


Fig. 3. Schematic description of the hybrid recursive Bayesian filtering algorithm for SOC estimation.

The algorithm of the recursive Bayesian filter (RBF) is formulated as follows:

- Initialization: The distribution of system state is:

$$p(x_0 | \mathbb{Y}_0) = \text{unif}(a, b) \quad (12)$$

where a and b are respectively the lower and upper limits of the system state x .

- State prediction for $k \geq 1$

$$p(x_k | \mathbb{Y}_{1:k-1}; \{\mathbb{U}_{1:k-1}, \Psi_{1:k}\}) = \int p(x_k | x_{k-1}; \Psi_k) \times p(x_{k-1} | \mathbb{Y}_{1:k-1}; \{\mathbb{U}_{1:k-1}, \Psi_{1:k-1}\}) dx_{k-1} \quad (13)$$

- State update for $k \geq 1$

The measurement update stage for $k \geq 1$ is formulated as:

$$p(x_k | \mathbb{Y}_{1:k}; \{\mathbb{U}_{1:k}, \Psi_{1:k}\}) = \frac{p(\mathbb{Y}_k | x_k; \mathbb{U}_k) p(x_k | \mathbb{Y}_{1:k-1}; \{\mathbb{U}_{1:k-1}, \Psi_{1:k}\})}{p(\mathbb{Y}_k | \mathbb{Y}_{1:k-1}; \{\mathbb{U}_{1:k}, \Psi_{1:k}\})} \quad (14)$$

where

$$p(\mathbb{Y}_k | \mathbb{Y}_{1:k-1}; \{\mathbb{U}_{1:k}, \Psi_{1:k}\}) = \int p(\mathbb{Y}_k | x_k; \mathbb{U}_k) p(x_k | \mathbb{Y}_{1:k-1}; \{\mathbb{U}_{1:k-1}, \Psi_{1:k}\}) dx_k \quad (15)$$

The optimal state estimation is achieved by different criterion. Examples are:

- Minimum mean-square error (MMSE)

$$\hat{x}_{k|k}^{\text{MMSE}} = E[x_k | \mathbb{Y}_{1:k}] = \int x_k p(x_k | \mathbb{Y}_{1:k}; \{\mathbb{U}_{1:k}, \Psi_{1:k}\}) dx_k \quad (16)$$

- Maximum a posteriori (MAP)

$$\hat{x}_{k|k}^{\text{MAP}} = \arg \max_{x_k} p(x_k | \mathbb{Y}_{1:k}; \{\mathbb{U}_{1:k}, \Psi_{1:k}\}) \quad (17)$$

Remark 3.1. The measurement model possibly involves a non-linear nonparametric discrete-valued conditional probability

distribution. Therefore, instead of using the commonly used Kalman filter, the general recursive Bayesian filter (RBF) might be the statistically correct way to compute the posterior distribution of the unobserved state [30]. Although the resulting computational cost would be larger to implement this RBF, there are benefits from using such an analytical measurement model as stated at the end of Subsection 3.1. Performance comparison of the proposed method and the conventional Kalman filter with a linear (or linearized) model with additive Gaussian measurement noise is recommended as a topic of future research as stated in Section 5.

4. Experimental validation of the SOC estimation algorithm

This section validates the proposed sequential estimation algorithm, which is a synergistic combination of both dynamic data-driven and model-based recursive analyses, on an ensemble of experimental data that have been collected from a commercial-scale lead-acid battery.

4.1. Experimental platform and data collection

This section addresses validation of the proposed SOC estimation method with the available experimental data that were used by Shen and Rahn in their recent publications [31,32]. Therefore, the experimental platform and data acquisition are not described in details in this paper. However, for the sake of completeness, a very brief description of the experimental platform and the data collection process is presented below.

The fresh (12 V AGM VRLA with 56 A h capacity) lead-acid battery, which has been used in the experiments, was charged and discharged according to the specified input (current) profiles at the room temperature. An ensemble of synchronized time-series of the input charge/discharge current and output voltage responses has been collected at the sampling frequency of 1 Hz. Fig. 4 exhibits typical profiles of long-term experimental data, namely, input current, output voltage, and battery SOC, and Fig. 5 presents short-term profiles of the corresponding experimental data. The input data set consists of fluctuating (with repeated ‘‘Hotel-Pulses’’

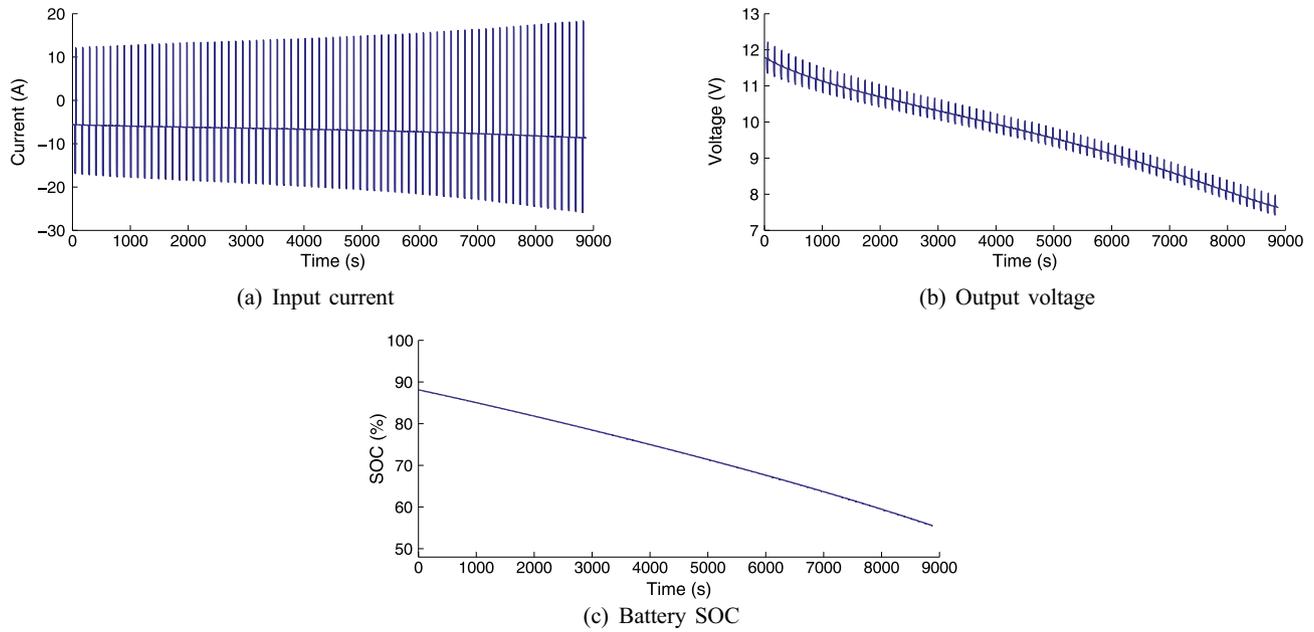


Fig. 4. Typical profiles of long-term data: input current, output voltage, and battery SOC.

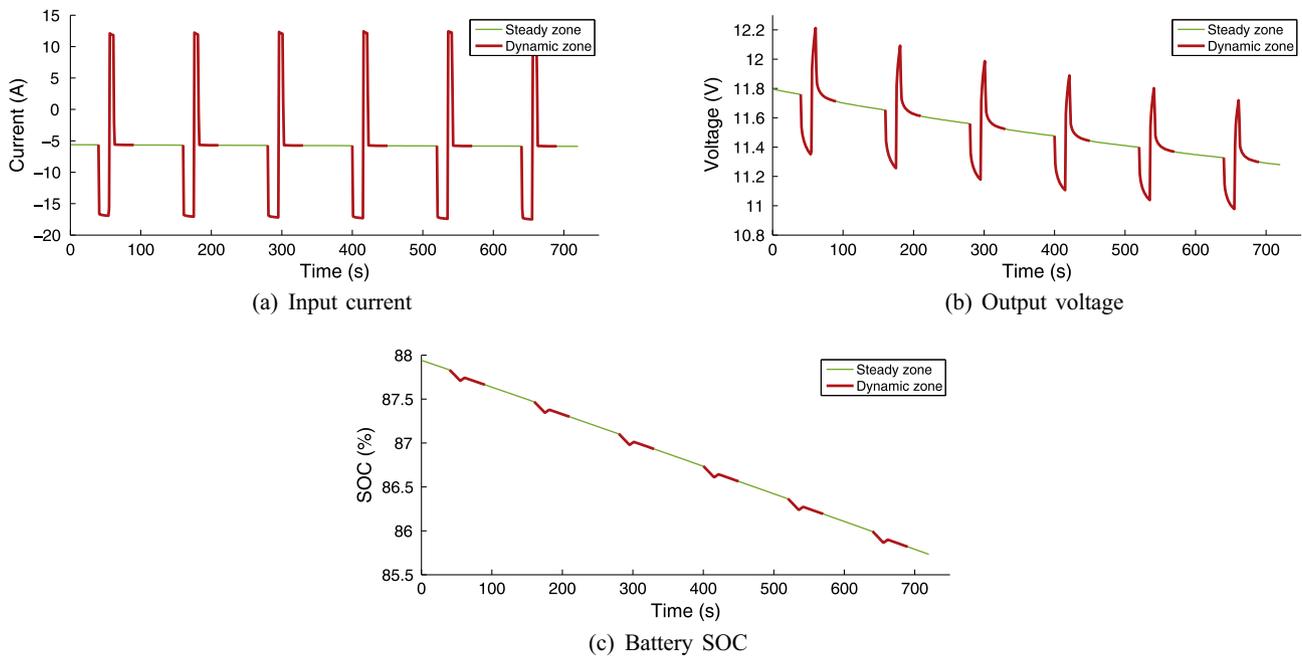


Fig. 5. Examples of dynamic zone and steady-state zone in the experimental data.

cycles) discharging/charging current, where each individual “Hotel-Pulse” cycle (with a duration of ~ 120 s) consists of a “hotel” load (i.e., relatively steady discharge due to “hotel” needs like lighting and other electrical appliances) and a discharge pulse followed by a charge (i.e., regeneration) pulse. The amplitude of the “hotel” load and the discharging & charging pulses are fluctuating in the experiment, which made each cycle different from others. This pattern of input cycles largely simulates a real-time working condition for an electric vehicle.

In Fig. 5, the highlighted zones are selected as analysis windows for the SOC estimation algorithm. The data length for one dynamic

zone (i.e., analysis window) is 50 on the fast time-scale in the unit of seconds.

4.2. Results and discussion

There are, in total, 25 experiments that have been conducted for this research, where the nominal battery maximum capacity is assumed to be the same for all experiment data.

In order to evaluate the robustness of the proposed approach, cross validation is performed in this paper. One round of cross-validation involves partitioning the data set randomly into comple-

mentary subsets, training the algorithm on one subset (training set), and testing the algorithm on the other subset (testing set). The size of the training and testing sets are taken to be in the ratio of 4:1. To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the estimation results are averaged over the rounds. In this paper the estimation results are obtained by averaging over 100 random partitions of the data set into training and testing sets.

The chosen values of the parameters in the data-driven measurement model and system state-space model are as follows:

- Sampling time for the fast time scale $\Delta t = 1$ s ; slow-scale time window length $n = 50$, i.e., consisting of 50 consecutive fast-scale time instants.
- The parameter for (Gaussian) radial basis function kernel:

$$\gamma = \frac{1}{2 * (0.03)^2} \tag{18}$$

- Maximum capacity of the battery during the k th window = initial battery capacity

$$C_k = C_{initial} \tag{19}$$

- Standard deviation for process noise $w_k = 10\%$ of the ratio of DOD Ψ_k and capacity C_k :

$$v_k = 0.1 * \frac{\Psi_k}{C_k} \tag{20}$$

In this section, the performance of hybrid recursive Bayesian filtering (RBF) is compared with that of data-driven estimation via $\times D$ -Markov modeling. The minimum mean-square error (MMSE) estimation is chosen as the criterion for the RBF at each step, while

the maximum a posteriori (MAP) estimation is used for the data-driven estimation:

- SOC estimation by hybrid recursive Bayesian filtering:

$$\hat{x}_{k|k}^{MMSE} = E[x_k | \mathbb{Y}_{1:k}] \tag{21}$$

- SOC estimation by solo data-driven $\times D$ -Markov modeling:

$$\hat{x}_{k|k}^{MAP} = \arg \max_{x_k} p(x_k | \mathbb{Y}_{1:k}; \{\cup_{1:k}, \Psi_{1:k}\}) \tag{22}$$

Fig. 6 depicts the performance of the hybrid SOC estimation algorithm for different alphabet sizes: $|\Sigma| = 4, 6$ and 8. For $|\Sigma| = 4$, the $\times D$ -Markov morph-matrix features are unable to capture the dynamic characteristics of the battery in the high range (e.g., $\sim 75\text{--}80\%$) and low range (e.g., $\sim 60\text{--}65\%$) of SOC. The execution time of the recursive estimation algorithm for convergence is in the order of one hour ≈ 30 analysis windows. However, as the alphabet size is increased to $|\Sigma| = 6$, the $\times D$ -Markov morph-matrix features improve their distinguishing capability at a low SOC range (e.g., $\sim 60\text{--}65\%$), while the convergence time is significantly reduced to ~ 20 analysis windows on the average. The rationale for this performance enhancement is explained below.

As the alphabet size is modestly increased from $|\Sigma| = 4$ to $|\Sigma| = 6$ and then to $|\Sigma| = 8$, the resulting $\times D$ -Markov machines capture more information of the dynamics of the time series; consequently, the corresponding morph-matrix features obtained under different operating conditions are more distinctive (i.e., well separated in the feature space); the effects are observed in Fig. 6. However, due to the finite length of the time series (i.e., finite length of the analysis window) under consideration, $|\Sigma|$ cannot

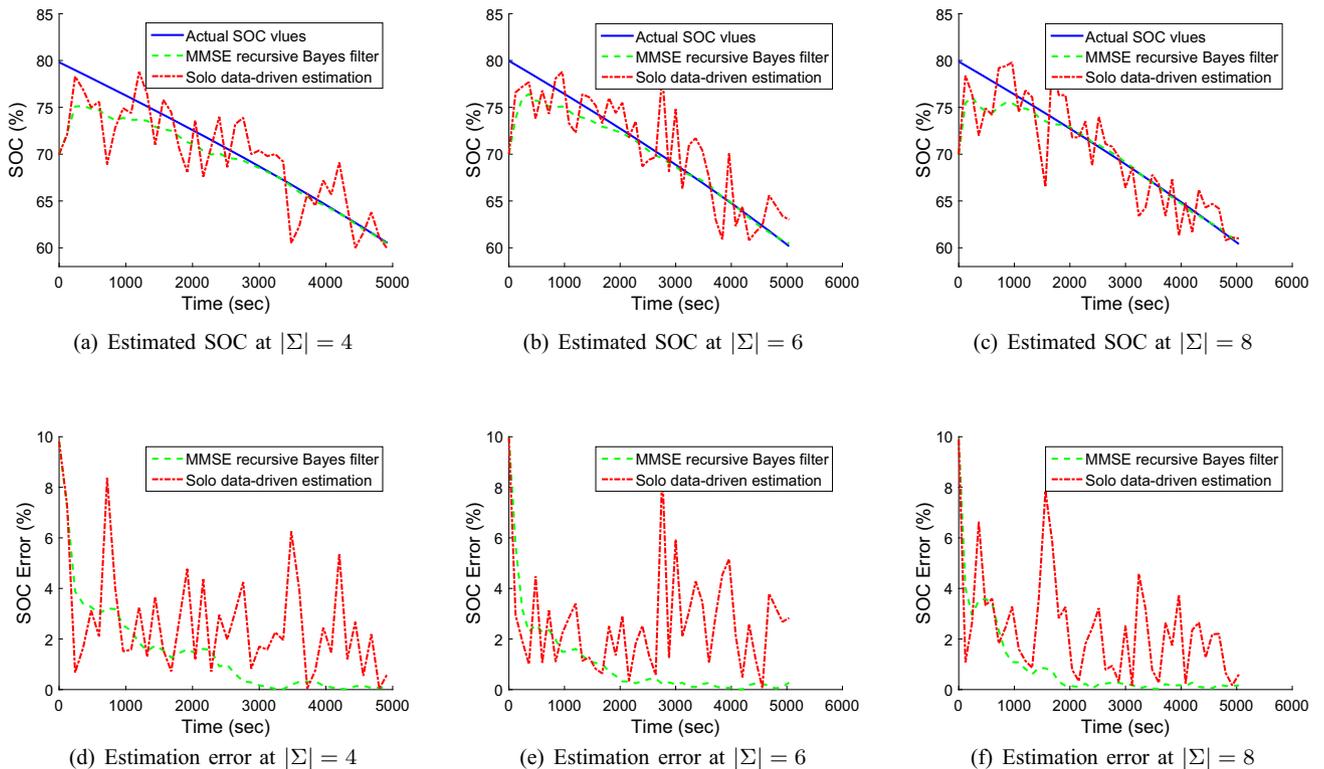


Fig. 6. Examples of recursive SOC estimation for different alphabet size $|\Sigma|$.

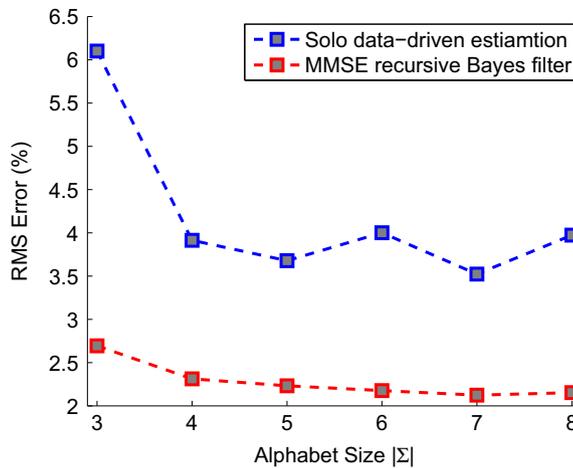


Fig. 7. Root-mean-square (RMS) error of SOC estimation for different alphabet size $|\Sigma|$.

Table 1
SOC estimation error comparison.

Method	RMS error (%)
Linear Model 1 in [32]	4.93
Linear Model 2 in [32]	5.04
Switched Model 3 in [32]	0.35
Switched Model 4 in [32]	0.20
Averaged Model 5 in [32]	10.00
The proposed method with $ \Sigma = 7$	2.08

be increased beyond a certain limit as explained by Wen et al. [33] and experimentally observed by Mukherjee and Ray [23].

Fig. 7 depicts the root-mean-square (RMS) error of SOC estimation for the data-driven and the hybrid approaches. The proposed hybrid approach performs better than the standalone data driven approach for all alphabet sizes, as seen in Fig. 7. Based on a mathematical model [34] of an electrochemical lead-acid battery cell developed earlier, Shen and Rahn [32] proposed five different models including two switched linear models that were linearized to produce charge, discharge and averaged models. Table 1 compares the performance of the proposed hybrid method of SOC estimation with that of the five solely model-based methods [32]. With the exception of both switched linear models (i.e., Model 3 and Model 4), the proposed method outperforms the (remaining three) regular linear models developed from an electrochemical cell model. It is expected that the proposed hybrid method, if applied on switched models instead of a single model in Eq. (9), would exceed the performance of Model 3 and Model 4 in [32]; this task is suggested as a topic of future research in Section 5.

The above discussion on comparison with other models evinces that the proposed hybrid method has a potential of successful real-life applications, where this method would provide competitive performance with less computational resources. For example, Li et al. [35] have recently reported the computation cost (i.e., execution time and memory requirements) of a dynamic data-driven method for battery SOC estimation.

5. Summary, conclusions and future work

This paper presents a hybrid (i.e., a combination of dynamic data-driven and model-based recursive) method for real-time estimation of a critical battery parameter, namely, state of charge (SOC). The underlying concept is built upon a measurement model and prior history of estimation, which rely on synchronized time

series pair of (repeated-pattern) fluctuating input (i.e., charging/discharging current) and output (i.e., battery voltage) data. The SOC estimation algorithm has been validated on experimental data of a commercial-scale lead-acid battery.

The pertinent conclusions drawn from the work reported in this paper are delineated below.

- (1) The proposed algorithm of dynamic data-driven and model-based recursive method of SOC estimation advances the state of the art of data-driven methods for both SOC estimation [14] and SOH estimation [35] as an alternative to the commonly used techniques (e.g., equivalent circuit model [31]).
- (2) The proposed method has two major contributions to the current state of the art of battery technology:
 - Development of a SOC measurement model via statistical regression, which is based on symbolization of the (synchronized) input and output time series.
 - Real-time estimation of SOC with significantly reduced computational complexity relative to the methods that rely on physics and electro-chemistry-based dynamic models of the battery system.

While there are several issues that need to be resolved by further theoretical and experimental research before implementation of the proposed method in real-life applications, the authors suggest the following topics of future research:

- *Validation of the proposed hybrid (i.e., combined dynamic data-driven and model-based recursive) method of SOC estimation for:*
 - Performance evaluation of the proposed method by using switched models (e.g., Model 3 and Model 4 in [32]) instead of a single model (e.g., in Eq. (9)).
 - Performance comparison of the proposed method with other recursive estimation algorithms (e.g., different types of Kalman filtering and particle filtering).
 - Validation of the proposed method on other types (e.g., Lithium-ion and Ni-MH) of battery with different input current profiles, such as Federal Urban Driving Schedule (FUDS) and US06 [36].
- *Classification of input profile for efficient battery management:* For example, in electric vehicles where Lithium-ion batteries are largely prevalent, the current input depends on the user's driving style, and the PFSA model of input current could classify the driving style into different categories.
- *Theoretical and experimental research to investigate specific advantages and disadvantages of the proposed method in comparison with existing methods (e.g., [6–8]):* One example of an existing is an equivalent circuit model with one RC pair. Another example is a conventional Kalman filter having a linear (or linearized) model and additive Gaussian measurement noise.
- *Development of a stopping rule for sequential estimation:* After reaching the stopping point, the estimation accuracy should be maintained at a specified level of SOC (e.g., by filtering out the potential outliers in the future data).

References

- [1] Wang Y, Zhang C, Chen Z. A method for state-of-charge estimation of li-ion batteries based on multi-model switching strategy. *Appl Energy* 2015;137:427–34.

- [2] Dong G, Wei J, Zhang C, Chen Z. Online state of charge estimation and open circuit voltage hysteresis modeling of lifepo4 battery using invariant imbedding method. *Appl Energy* 2016;162:163–71.
- [3] Hu X, Sun F, Zou Y. Comparison between two model-based algorithms for li-ion battery SOC estimation in electric vehicles. *Simul Model Pract Theory* 2013;34:1–11.
- [4] Tanim TR, Rahn CD, Wang C-Y. A temperature dependent, single particle, lithium ion cell model including electrolyte diffusion. *J Dyn Syst Meas Control* 2015;137(1):011005.
- [5] Plett GL. Extended Kalman filtering for battery management systems of LiPb-based HEV battery packs: part 3 – state and parameter estimation. *J Power Sources* 2004;134(2):277–92.
- [6] Xiong R, He H, Sun F, Zhao K. Evaluation on state of charge estimation of batteries with adaptive extended Kalman filter by experiment approach. *IEEE Trans Veh Technol* 2013;62(1):108–17.
- [7] Sepasi S, Ghorbani R, Liaw BY. Soc estimation for aged lithium-ion batteries using model adaptive extended Kalman filter. In: 2013 IEEE Transportation Electrification Conference and Expo (ITEC). IEEE; 2013. p. 1–6.
- [8] Sepasi S, Ghorbani R, Liaw BY. Improved extended Kalman filter for state of charge estimation of battery pack. *J Power Sources* 2014;255:368–76.
- [9] He W, Williard N, Chen C, Pecht M. State of charge estimation for electric vehicle batteries using unscented Kalman filtering. *Microelectron Reliab* 2013;53(6):840–7.
- [10] Sun F, Hu X, Zou Y, Li S. Adaptive unscented Kalman filtering for state of charge estimation of a lithium-ion battery for electric vehicles. *Energy* 2011;36(5):3531–40.
- [11] Liu X, Chen Z, Zhang C, Wu J. A novel temperature-compensated model for power li-ion batteries with dual-particle-filter state of charge estimation. *Appl Energy* 2014;123:263–72.
- [12] Darema F. Dynamic data driven applications systems: new capabilities for application simulations and measurements. In: 5th International Conference on Computational Science – ICCS 2005, (Atlanta, GA; United States). p. 610–5.
- [13] Shi Q-S, Zhang C-H, Cui N-X. Estimation of battery state-of-charge using v-support vector regression algorithm. *Int J Automotive Technol* 2008;9(6):759–64.
- [14] Li Y, Chattopadhyay P, Ray A, Rahn CD. Identification of the battery state-of-health parameter from input-output pairs of time series data. *J Power Sources* 2014;268:758–64.
- [15] Hu C, Jain G, Zhang P, Schmidt C, Gomadam P, Gorka T. Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery. *Appl Energy* 2014;129:49–55.
- [16] Anton A, Carlos J, Garcia Nieto PJ, Blanco Viejo C, Vilan V, Antonio J. Support vector machines used to estimate the battery state of charge. *IEEE Trans Power Electron* 2013;28(12):5919–26.
- [17] Kang L, Zhao X, Ma J. A new neural network model for the state-of-charge estimation in the battery degradation process. *Appl Energy* 2014;121:20–7.
- [18] Charkhgard M, Farrokhi M. State-of-charge estimation for lithium-ion batteries using neural networks and ekf. *IEEE Trans Ind Electron* 2010;57(12):4178–87.
- [19] Xu L, Wang J, Chen Q. Kalman filtering state of charge estimation for battery management system based on a stochastic fuzzy neural network battery model. *Energy Convers Manage* 2012;53(1):33–9.
- [20] He W, Williard N, Chen C, Pecht M. State of charge estimation for li-ion batteries using neural network modeling and unscented Kalman filter-based error cancellation. *Int J Electr Power Energy Syst* 2014;62:783–91.
- [21] Li Y, Shen Z, Ray A, Rahn CD. Real-time estimation of lead-acid battery parameters: a dynamic data-driven approach. *J Power Sources* 2014;268:758–64.
- [22] Ray A. Symbolic dynamic analysis of complex systems for anomaly detection. *Signal Processing* 2004;84:1115–30.
- [23] Mukherjee K, Ray A. State splitting and merging in probabilistic finite state automata for signal representation and analysis. *Signal Processing* 2014;104:105–19.
- [24] Sarkar S, Jha DK, Ray A, Li Y. Dynamic data-driven symbolic causal modeling for battery performance & health monitoring. In: 2015 18th international conference on information fusion (Fusion). IEEE; 2015. p. 1395–402.
- [25] Chattopadhyay P, Li Y, Ray A. Real-time identification of state-of-charge in battery systems: dynamic data-driven estimation with limited window length. In: American Control Conference, Boston, MA, USA.
- [26] Rajagopalan V, Ray A. Symbolic time series analysis via wavelet-based partitioning. *Signal Processing* 2006;11:3309–20.
- [27] Vidal E, Thollard F, De La Higuera C, Casacuberta F, Carrasco RC. Probabilistic finite-state machines-part i. *IEEE Trans Pattern Anal Mach Intell* 2005;27(7):1013–25.
- [28] Rahn CD, Wang C-Y. *Battery systems engineering*. John Wiley; 2012.
- [29] Beim Graben P. Estimating and improving the signal-to-noise ratio of time series by symbolic dynamics. *Phys Rev E* 2001;64(5):051104.
- [30] Särkkä S. *Bayesian filtering and smoothing*, vol. 3. Cambridge University Press; 2013.
- [31] Shen Z, Rahn C. Physics-based model of a valve-regulated lead-acid battery and an equivalent circuit. *J Dyn Syst Meas Contr* 2013;135:041011-1–1-7.
- [32] Shen Z, Rahn C. Model-based state-of-charge estimation for a valve-regulated lead-acid battery using matrix inequalities. *J Dyn Syst Meas Contr* 2013;135:041015-1–5-8.
- [33] Wen Y, Mukherjee K, Ray A. Adaptive pattern classification for symbolic dynamic systems. *Signal Processing* 2013;93:252–60.
- [34] Shen Z, Gou J, Rahn CD, Wang C-Y. Ritz model of a lead-acid battery with application to electric locomotives. In: ASME 2011 dynamic systems and control conference and bath/ASME symposium on fluid power and motion control. American Society of Mechanical Engineers; 2011. p. 713–20.
- [35] Li Y, Chattopadhyay P, Ray A. Dynamic data-driven identification of battery state-of-charge via symbolic analysis of input-output pairs. *J Appl Energy* 2015;155:778–90.
- [36] Johnson VH, Pesaran AA, Sack T, America S. *Temperature-dependent battery models for high-power lithium-ion batteries*. Golden, CO, USA: National Renewable Energy Laboratory; 2001.