Home Appliance Load Modeling From Aggregated Smart Meter Data

Zhenyu Guo, Z. Jane Wang, *Senior Member, IEEE*, and Ali Kashani

***Abstract*—With recent developments in the infrastructure of smart meters and smart grid, more electric power data is available and allows real-time easy data access. Modeling individual home appliance loads is important for tasks such as non-intrusive load disaggregation, load forecasting, and demand response support. Previous methods usually require sub-metering individual appliances in a home separately to determine the appliance models, which may not be practical, since we may only be able to observe aggregated real power signals for the entire-home through smart meters deployed in the field. In this paper, we propose a model, named Explicit-Duration Hidden Markov Model with differential observations (EDHMM-diff), for detecting and estimating individual home appliance loads from aggregated power signals collected by ordinary smart meters. Experiments on synthetic data and real data demonstrate that the EDHMM-diff model and the specialized forward-backward algorithm can effectively model major home appliance loads.**

***Index Terms*—Disaggregation, explicit duration hidden Markov model (HMM), forward-backward, load modeling.**

I. Introduction

As part of the smart grid deployment, smart meters can provide more energy consumption information than we could imagine before in a near real-time way. With increasing installations of smart meters in more countries, such as Australia, Canada, Italy, Japan, United States, etc., massive amount of residential electric energy consumption data has been collected and stored. Although current advanced infrastructures of smart grid could provide full potentials for advanced services, insightful analysis and modeling based on such big data is still in its early stage. Exploration of such valuable data emerges as a popular research direction both in academia and industry, and conventional services, such as load disaggregation (LD), load forecasting (LF) and demand response (DR) support, are brought back to attention. Modeling the home appliance loads plays an important role for these applications, since it is the first step for understanding the electric consumption data. In this paper, we focus on modeling home appliance loads under a general assumption, where only *aggregated* real-time power data is observed by ordinary smart meters already deployed, with a *low* data sampling rate.

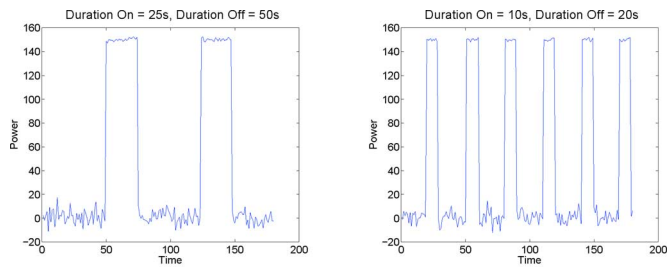


Fig. 1. Signals shown in (a) and (b) are two signals with the same emission probabilities but different state duration distributions. X axis is abstract index without unit for illustration purpose. A conventional HMM model cannot capture the difference between them since the state durations are not considered.

Starting from Hart [1], power consumption signatures [2], [3] are used to describe the behaviors of home appliances, which include information such as time of use, on-and-off durations and patterns, power demands, etc. Some signatures also encode the **transient** properties of the current and voltage signals of appliances when they are turned on or off, while some signatures mainly focus on **stable** properties of the power signals. Most smart meters installed in the field measure and transmit the real power signals of residential users at a relatively low frequency ($1\ \text{Hz} \backsim (1)/(900)\ \text{Hz}$). Therefore, the low sampling rate makes stable signature a more suitable choice for home appliance load modeling. Most home appliances work at one or several fixed power demands, which can be characterized by finite discrete states. In addition, one power reading at present is independent from early readings in the past. Therefore, hidden Markov model (HMM) [4] seems a good choice and is widely used to model home appliances to extract stable information.

However, the conventional HMM can not model the duration of each state, which is important for estimating the electric energy consumption of a home appliance. The states' durations and switching patterns are also crucial for describing appliances, and could help increase the accuracy of detection and estimation. In Fig. 1 we illustrate a confusion caused by the lack of duration modeling of conventional HMM on power signals.

In many realistic situations, the aggregated power signal is the only data collected from one household, and the estimation of individual appliance models can only be done based on the aggregated signals. Since several different appliances could be turned on within the same period, the real power signal of one

Manuscript received September 16, 2013; revised February 14, 2014 and April 01, 2014; accepted May 16, 2014. Date of publication June 05, 2014; date of current version December 18, 2014. Paper no. TPWRS-01143-2013.

Z. Guo is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, and also with Energy Aware Technology Inc., Vancouver, BC V6A 3X3, Canada (e-mail: zhenyug@ece.ubc.ca).

Z. J. Wang is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: zjanew@ece.ubc.ca).

A. Kashani is with Energy Aware Technology Inc., Vancouver, BC V6A 3X3, Canada (e-mail: ali.kashani@energy-aware.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2014.2327041

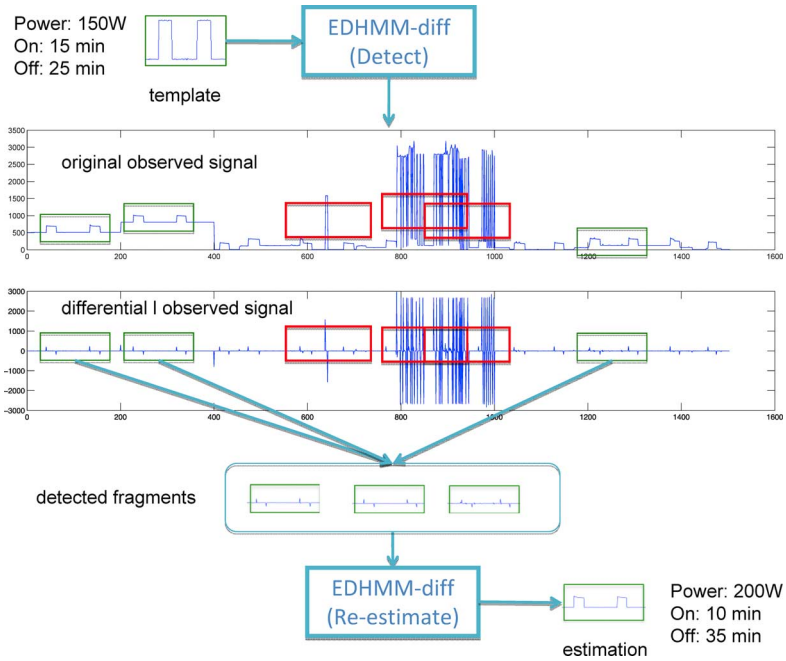


Fig. 2. Illustration of the "detect and re-estimate" approach for the EDHMM-diff model. Y axis is power in Watts, and X axis is time in a unit of a half minute. From top to bottom: A template is used to detect signal clips generated by a certain AOI; Red boxes indicate rejected signal fragments and green boxes indicate accepted fragments. The accepted (green boxes) fragments are concatenated for estimating the true model using the EDHMM-diff estimation algorithm in Table I, and a final estimation of the AOI is obtained.

TABLE I
PROPOSED FORWARD-BACKWARD ALGORITHM FOR THE EDHMM-DIFF MODEL

Algorithm 1 The Forward-backward Algorithm for the EDHMM-diff

Set $iter = 1$, $lkh_{prev} = 0$.
While $iter <$ **maxIter and** $\triangle lkh <$ *threshold*
1: Initialize $\alpha_{t|t-1}$ at $t = 2$ according to Eq.(11).
2: **Forward Induction**: for $t = 3, \ldots, T$
Compute and store $\alpha_{t|t-1}$ using Eq.(13), ε_t using Eq.(14), S_t using Eq.(15), and γ_t^{-1} using Eq.(16).
3: Initialize β_t at $t = T$ according to Eq.(12).
4: **Backward Induction**: for $t = T-1, \ldots, 2$
Compute β_t in Eq.(17), ε_t^* in Eq.(18), and S_t^* in Eq.(19).
Compute and store $\mathcal{T}_{t|T}$ in Eq.(21), $\mathcal{D}_{t|T}$ in Eq.(23), and $\gamma_{t|T}$ in Eq.(24).
5: Update the parameters:
$\hat{A}_{ij}$ in Eq.(25), $P_j(d)$ in Eq.(26), and $\hat{\mu}_{i,j}$ in Eq.(27).
6: Compute the likelihood:
lkh_{curr} in Eq.(20), $\triangle lkh = |lkh_{curr} - lkh_{prev}|$.
7: Set $iter = iter + 1$, $lkh_{prev} = lkh_{curr}$.
end While

appliance could be "lifted" or overlapped with power signals of other appliances that are turned on during the same time. We call such phenomenon as "aggregating effect". In Fig. 2, an aggregated power signal is shown as the "original observed signal", where the green boxes indicate real power signals generated by the same refrigerator in the house. Although these underlying "refrigerator signals" are generated by the same appliance, the power signals observed in these boxes are somehow different due to the "aggregating effect". Since the emission probabilities used in conventional HMM are modeled directly on the observations, the conventional HMM cannot handle the "aggregating effect", which generates different observations even for the same state of the same device.

To overcome the problems mentioned above, in this paper, we propose an Explicit Duration Hidden Markov Model with differential observations (EDHMM-diff), along with a specialized forward-backward algorithm for the inference and estimation of EDHMM-diff model. In addition to the information that can be learned in the conventional HMM, EDHMM-diff can estimate the model of individual appliances based on the aggregated power signal with state durations. Furthermore, in most cases, only a few appliances are of interest to the utilities or users, which we refer to as "Appliance of Interest" (AOI). Accordingly, we propose a "detect and re-estimate" approach, which uses a predefined template to detect the best fragments for estimating one AOI, and then re-estimates the template model using these detected fragments. In real-world applications, given the AOIs, we could "detect and re-estimate" them separately. The framework of the proposed method is shown in Fig. 2.

LD is one of the most important applications of smart grid data analysis, which aims to figure out what appliances are used in a home as well as their individual energy consumptions, by

only observing the aggregated electric consumption data for the entire-home. factorial hidden Markov mode (FHMM) based modeling [5]–[7] is one recent promising direction for LD research, which shows satisfiable disaggregation results on real data. However, all previous FHMM methods require correctly estimated models of individual HMM chains, where manually efforts are unavoidable [5], [7]. As one example, our proposed method could be used as an automatic step of estimation of individual HMMs for other FHMM based methods. In addition, our proposed method is a general approach for situations where individual device models are required. For instance, our method could be used to estimate the power efficiency of the refrigerator used in a monitored house, to remind the user to replace the refrigerator with a more efficient ones to save money. The rest of the paper is organized as following, in Section II, we will formulate the research problem and propose the EDHMM-diff algorithm as a solution. We will conduct experiments both on synthetic data and real data in Section III. At last, we conclude the paper in Section IV.

II. Proposed Method

In this section, we will describe the EDHMM-diff model for appliance load modeling, and propose a corresponding specialized algorithm for inference and estimation.

A. Explicit-Duration Hidden Markov Model With Differential Observations (EDHMM-Diff)

Under the current implementation of smart meters, the observation is a sequence of aggregated real power readings, which can be denoted as $\mathbf{o} = (o_1, o_2, \ldots, o_t, \ldots, o_T)$. Suppose there are N appliances in the home, then the power signal generated by the lth appliance at time t is denoted as $x_t^{\{l\}}$, therefore we have $o_t = \sum_{l=1}^{N} x_t^{\{l\}}$. In previous works [5]–[7], a conventional HMM is used to model each individual appliance. Each appliance is assumed to work at finite discrete states, i.e., two (on and off) states for a light; and cooking, warming, and off 3 states for a cooker. For appliance l, the lth HMM, $\text{HMM}^{\{l\}}$, can be described as

$$\begin{aligned} &P\left(q_1^{\{l\}} = k\right) = \pi_k^{\{l\}} \\ &P\left(q_t^{\{l\}} = j \middle| q_{t-1}^{\{l\}} = i\right) = A_{ij}^{\{l\}} \\ &x_t^{\{l\}} \middle| q_t^{\{l\}} = k \sim \mathcal{N}\left(\mu_k^{\{l\}}, \left(\sigma_k^{\{l\}}\right)^2\right) \end{aligned} \tag{1}$$

where $q_t^{\{l\}}$ is the hidden state variable for the lth HMM at time t, which takes a discrete value from a finite set $\{1, 2, \ldots M_l\}$. Also $\pi_k^{\{l\}}$ is the corresponding initial probability, $A_{ij}^{\{l\}}$ is the corresponding entry in the transition matrix $A^{\{l\}}$, and $\mu_k^{\{l\}}$ and $\sigma_k^{\{l\}}$ are the parameters for the Gaussian emission probability density for the ith HMM at state k. We denote the set of parameters of the ith HMM as $\Theta^{\{l\}}$, and the set of all parameters of all HMMs for N appliances as $\{\Theta^{\{l\}}\}_{l=1}^{N}$.

In this paper, we focus on how to estimate some $\Theta^{\{l\}}$'s from $\{\Theta^{\{l\}}\}_{l=1}^{N}$, corresponding to certain AOI's, based on the aggregated observations $\mathbf{o}$. As we pointed out in Section I, a conventional HMM cannot capture the information of the durations of individual states, which are important to identify individual appliances and to estimate their energy consumptions. To solve this problem, we introduce a probability distribution over the duration of each state in the conventional $\text{HMM}^{\{l\}}$, which can be described as

$$P_j^{\{l\}}(d) = P\left(\tau^{\{l\}} = d \middle| q^{\{l\}} = j\right) \tag{2}$$

where $\tau^{\{l\}}$ is the duration of state $q^{\{l\}}$ staying at j.

In addition, the "aggregating effect" described in Section I also causes a big trouble for evaluating and estimating the emission probability. As shown in Fig. 2, when the power signal of a refrigerator is "lifted" by a light's power signal by 500 W, the resulting refrigerator power at the "off-state" becomes 500 W and becomes 750 W at the "on-state", which leads to a small probability when fitting the data with the emission distribution of that refrigerator. To deal with this concern, we adapt an reasonable assumption, which is widely accepted in the load modeling and load disaggregation research area [5]–[8], that the probability for more than one appliance to change state is very low within a short period. In another word, we assume that at most one appliance changes state within a short period. Therefore, we could find segments of signals where only the lth appliance changes state, which is defined as Regions of Interest (ROIs) for appliance l. Let $\mathbf{y} = (y_1, y_2, \ldots, y_t, \ldots, y_T)$ denote the differential signal of the original $\mathbf{o}$, where $y_t = o_t - o_{t-1}$ with $o_0 = 0$. Let $\delta x_t^{\{l\}} = x_t^{\{l\}} - x_{t-1}^{\{l\}}$ denote the differential signal of appliance l. Now suppose from t_1 to t_m is the ROI of appliance l, then $(y_{t_1}, \ldots, y_{t_m}) = (\delta x_{t_1}^{\{l\}}, \ldots, \delta x_{t_m}^{\{l\}})$, since only appliance l changes state during this period. Therefore, we propose addressing the "aggregating effect" by modeling the differential observations in ROIs instead. By adding duration modeling and differential observations to the conventional HMM, we propose the Explicit Duration Hidden Markov Model with Differential observations (EDHMM-diff), which can be described by

$$\begin{aligned} b_{i,j}(y_t) &= P(y_t \,|\, q_{t-1} = i, q_t = j) \\ P_j(d) &= P(\tau = d \,|\, q = j) \\ A_{ij} &= P(q_t = j \,|\, q_{t-1} = i) \\ \pi_i &= P(q_1 = i) \end{aligned} \tag{3}$$

where we omit the $\{l\}$ notation and use it as a general model for a particular appliance. $b_{i,j}(y_t)$ is the emission probability of the differential observation given two adjacent hidden states q_{t-1} and q_t. Other variables retain the same meanings as previous. A graphical illustration is shown in Fig. 3. Although it seems that the only difference between our model and Explicit Duration model [9] is that we use differential signals as observations, our contribution is not trivial, since the proposed EDHMM-diff model is a second order model (observation depends on two states) and a specialized forward-backward algorithm needs to be invented for inference and estimation, which is presented in Section II-B. A practical solution for inference and estimation plays an important role in any HMM based non-parametric methods.

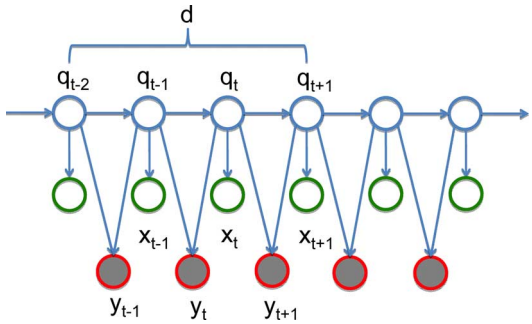


Fig. 3. Graphical illustration of the proposed EDHMM-diff model, where $y_t = o_t - o_{t-1}$, q_t is the hidden state at time t, and d is the duration of the hidden state.

In our implementation, we assume Gaussian distributions for the emission probability and the duration, which are

$$\begin{aligned} y_t \,|\, q_t = j, q_{t-1} = i &\sim \mathcal{N}(\delta\mu_{ij}, (\delta\sigma_{ij})^2) \\ d \,|\, q = k &\sim \mathcal{N}(d\mu_k, (d\sigma_k)^2) \end{aligned} \tag{4}$$

where $\delta\mu_{ij} = \mu_j - \mu_i$, and ${\delta\sigma_{ij}}^2 = {\sigma_i}^2 + {\sigma_j}^2$. Also $d\mu_k$ and $d\sigma_k$ are the mean and variance of the duration of a state staying at k. In the practical implementation, we actually use a discrete Gaussian distribution (sampled from a continuous Gaussian) for the duration. However we still use the same notations in the rest of the paper for consistency.

For convenience of expression, we call the set of parameters of an EDHMM-diff as Θ, which contains $\{\delta\mu, \delta\sigma, P(d), A, \pi\}$. Among these parameters, $\delta\mu$ and $P(d)$ are the two we are particularly interested in for home appliance load modeling. In following sections, we will present a practical algorithm for estimating Θ.

B. Estimation

To apply EDHMM-diff in a real-world application for a given signal $Y_{t_1:t_2} = \{y_{t_1}, y_{t_1+1}, \ldots, y_{t_2}\}$, we need to compute the likelihood $P(Y_{t_1:t_2} \,|\, \Theta)$, and efficiently estimate the parameters by maximizing the likelihood. Similar to the conventional HMM, straightforwardly calculating the likelihood is computationally infeasible and suffers from the floating-point "underflow" problem. To provide efficient estimation and inference for EDHMM-diff, here we propose a specialized forward-backward procedure, inspired by [6], [9], and [10]. The pseudo code is given in Table I.

1) Definitions of Variables: In addition to the variables defined in (3), we also define some auxiliary variables for the forward-backward algorithm. A **forward variable** is defined as

$$\alpha_{t\,|\,\lambda}(i,j,d) \overset{\text{def}}{=} P(q_{t-1} = i, q_t = j, \tau_t = d \,|\, Y_{2:\lambda}) \tag{5}$$

where λ can be $t-1, t$, or T, which corresponds to the "predicted", "filtered", or "smoothed" probability of the triplet (q_{t-1}, q_t, τ_t). The calculation of the auxiliary variable is usually a iterative process involving multiplication of a large number of probability values (small positive numbers less than 1), so that the result will be a very small float number that cannot be handled by a computer. Such problem is called "arithmetic underflow", or "underflow" for short. To overcome the "underflow" problem caused by multiplications of a large number of small probability values, we normalize the emission probability at every t as

$$b^*_{i,j}(y_t) \overset{\text{def}}{=} \frac{b_{i,j}(y_t)}{P(y_t \,|\, Y_{2:t-1})} \tag{6}$$

which approaches 1 when the fit of the observation to the model increases, and reaches 1 when the observation fits the model exactly. Such normalization can successfully avoid "underflow" and maintain other conditions required in the inference. For convenience, we denote the denominator probability as

$$\gamma_t^{-1} \overset{\text{def}}{=} P(y_t \,|\, Y_{2:t-1}) \tag{7}$$

which can be computed recursively. To clearly demonstrate the recursion, we define several other auxiliary variables

$$\begin{aligned} S_t(i,j) &\overset{\text{def}}{=} P(q_t = i, q_{t+1} = j, \tau_t = 1 \,|\, Y_{2:t}) \\ \varepsilon_t(i,j) &\overset{\text{def}}{=} P(q_{t-1} = i, q_t = j, \tau_t = 1 \,|\, Y_{2:t}). \end{aligned} \tag{8}$$

To calculate the smoothed probabilities, we define a **backward variable** which is a standard smoothed probability **normalized** by the predicted one as

$$\beta_t(i,j,d) \overset{\text{def}}{=} \frac{P(Y_{t:T} \,|\, q_{t-1} = i, q_t = j, \tau_t = d)}{P(Y_{t:T} \,|\, Y_{2:t-1})}. \tag{9}$$

To compute this backward variable recursively, we define another two auxiliary variables as

$$\begin{aligned} S^*_t(i,j) &\overset{\text{def}}{=} \frac{P(Y_{t:T} \,|\, q_{t-2} = i, q_{t-1} = j, \tau_{t-1} = 1)}{P(Y_{t:T} \,|\, Y_{2:t-1})} \\ \varepsilon^*_t(i,j) &\overset{\text{def}}{=} \frac{P(Y_{t:T} \,|\, q_{t-1} = i, q_t = j, \tau_{t-1} = 1)}{P(Y_{t:T} \,|\, Y_{2:t-1})}. \end{aligned} \tag{10}$$

So far we have defined necessary variables for the forward-backward induction. We now give details about the recursive *forward-backward* algorithm.

2) Forward-Backward Induction: Since the above auxiliary variables are defined in a recursive fashion, we need to initialize these variables at the beginning. The forward and backward variables are initialized as follows:

$$\alpha_{2\,|\,1}(i,j,d) = \pi_i A_{ij} P_j(d) \tag{11}$$

$$\beta_T(i,j,d) = b^*_{i,j}(y_T). \tag{12}$$

Then the forward variable and corresponding auxiliary variables can be updated as

$$\alpha_{t\,|\,t-1}(i,j,d) = \begin{cases} S_{t-1}(i,j) P_j(d), & \text{if } i \neq j, \\ \sum_k \alpha_{t-1\,|\,t-2}(k,j,d+1) b^*_{i,j}(y_{t-1}), & \text{if } i = j \end{cases} \tag{13}$$

$$\varepsilon_t(i,j) = \alpha_{t\,|\,t-1}(i,j,1) b^*_{i,j}(y_t) \tag{14}$$

$$S_t(i,j) = \sum_k \varepsilon_t(k,i) A_{ij} \tag{15}$$

$$\gamma_t^{-1} = \sum_{i,j,d} \alpha_{t\,|\,t-1}(i,j,d) b_{i,j}(y_t). \tag{16}$$

The backward variable and corresponding auxiliary variables are updated as

$$\beta_t(i,j,d) = \begin{cases} S^*_{t+1}(i,j)b^*_{i,j}(y_t), & \text{if } d=1, \\ \beta_{t+1}(j,j,d-1)b^*_{i,j}(y_t), & \text{if } d>1 \end{cases} \quad (17)$$

$$\varepsilon^*_t(i,j) = \sum_d \beta_t(i,j,d)P_j(d) \quad (18)$$

$$S^*_t(i,j) = \sum_k \varepsilon^*_t(j,k)A_{ij}. \quad (19)$$

So we have given the details for the forward-backward induction for the EDHMM-diff. We now present the re-estimation step in the next section.

3) Parameter Re-Estimation: After computing and storing these variables, we can calculate the likelihood and update the parameters of the EDHMM-diff model. The likelihood of the model $\mathcal{M}$ can be computed by

$$\begin{aligned} P(Y_{2:T} \mid \mathcal{M}) &= \prod_{t=3}^{T} P(y_t \mid Y_{2:t-1}) \\ &= \prod_{t=3}^{T} \gamma_t^{-1}. \end{aligned} \quad (20)$$

To update parameters of the EDHMM-diff model, we first define the following auxiliary variables:

$$\begin{aligned} \mathcal{T}_{t \mid T}(i,j) &= P(q_{t-1}=i, \tau_{t-1}=1, q_t=j \mid Y_{2:T}) \\ &= \varepsilon_{t-1}(i,j)\varepsilon^*_t(i,j) \end{aligned} \quad (21)$$

$$\begin{aligned} \mathcal{D}_{t \mid T}(i,j,d) &= P(q_{t-1}=i, q_t=j, \tau_{t-1}=1 \mid Y_{2:t-1}) \\ &\quad \times P(\tau_t = d \mid q_t = j) \\ &\quad \times \frac{P(Y_{t:T} \mid q_{t-1}=i, q_t=j, \tau_t=d)}{P(Y_{t:T} \mid Y_{2:t-1})} \\ &= S_{t-1}(i,j)P_j(d)\beta_t(i,j,d). \end{aligned} \quad (22)$$

$$\mathcal{D}_{t \mid T}(j,d) = \sum_i \mathcal{D}_{t \mid T}(i,j,d). \quad (23)$$

$$\begin{aligned} \gamma_{t \mid T}(i,j) &= P(q_{t-1}=i, q_t=j \mid Y_{2:T}) \\ &= \sum_d \alpha_{t \mid t-1}(i,j,d)\beta_t(i,j,d). \end{aligned} \quad (24)$$

Then the model parameters can be updated at each iteration as

$$\hat{A}_{ij} = \sum_{t=3}^{T} \mathcal{T}_{t \mid T}(i,j)/N_a \quad (25)$$

$$\hat{P}_j(d) = \sum_{t=3}^{T} \mathcal{D}_{t \mid T}(j,d)/N_p \quad (26)$$

$$\delta\hat{\mu}_{i,j} = \sum_{t=3}^{T} \gamma_{t \mid T} y_t / N_{\delta\mu} \quad (27)$$

$$\hat{\pi}_i = \frac{\sum_j \gamma_{2 \mid T}(i,j)}{N_\pi} \quad (28)$$

where $N_a, N_p, N_{\delta\mu}$, and N_π are normalization constants to make the resulting variables statistically valid. The forward-backward algorithm is summarized in Table I. Since we want to manually control the flexibility of the EDHMM-diff model for both detection and estimation purpose, we don't update the covariance $\delta\sigma$ of emission probability function. In practice, the $\delta\sigma$ is fixed to a number depending on the task.

C. Detection and Re-Estimation

As discussed in Section II-A, before estimating an appliance model through the EDHMM-diff, we need to detect the ROIs for this particular appliance. For this *detection* purpose, a prior model can be used as a template for this particular appliance. Although home appliances vary in terms of brands and models, the same type of appliances still share certain common characteristics in their power signals. For example, refrigerators usually work at a power demand of 70 W–200 W, with roughly a 20-min duration for each of the on/off states. Such consistent patterns are observed for appliances such as refrigerator, cloth dryer, cloth washer, dish washer, oven, etc., which are common AOIs for load disaggregation and forecasting. Therefore, the pattern of a certain AOI can be encoded as prior knowledge into a template EDHMM-diff model, which can be further used to detect ROIs for this particular appliance.

Let $\mathbf{o} = (o_1, o_2, \ldots, o_t, \ldots, o_T)$ denote the observations. Define $\text{window}_{t_1:t_2}(\mathbf{O})$ as the sequence $(o_{t_1}, o_{t_1+1}, \ldots, o_{t_2})$, a window of signals from time t_1 to t_2. Given a template model $\mathcal{M}_0$, the likelihood of $\mathcal{M}_0$ given the window of signals can be computed by using (20), denoted as $\mathcal{L}(\text{window}_{t_1:t_2})$. By sliding the window along the observations $\mathbf{o}$, we accept the windows whose likelihoods are above the threshold as training data for re-estimation. During the detection stage, only **steps 1, 2,** and **6** in Table I are run for each iteration.

After detecting the valid training data for each AOI, we update the template model according to the algorithm in Table I to get the final appliance model. The procedure described in this section is named "detect and re-estimate", which is illustrated in Fig. 2.

III. Experiment

In this section, we describe the experiments on both synthetic data and real data by applying the proposed method. On the synthetic data, we will estimate the EDHMM-diff model with the proposed forward-backward algorithm and report the estimation results. On the real data, we will apply the proposed "detect and re-estimate" procedure to learn models for individual appliances. The data used in Section III-B1 is from the REDD data set proposed by [8].

A. Simulation Study

To measure the inference and estimation effectiveness of the proposed forward-backward algorithm for the EDHMM-diff model, we conduct two sets of experiments.

In the first set of experiments, two time series with 500 data points are generated, which are shown in Fig. 4. The parameters to generate these two signals are listed in Table II. To simulate the "aggregating effect", we add a random DC value (with zero mean Gaussian noise) to each of these two signals. We use Gaussian distributions for both the emission probability and duration distribution for these two signals. It is reasonable to assume that most home appliances work with 0 power demand at the "off" state (although sometime a small value of power can

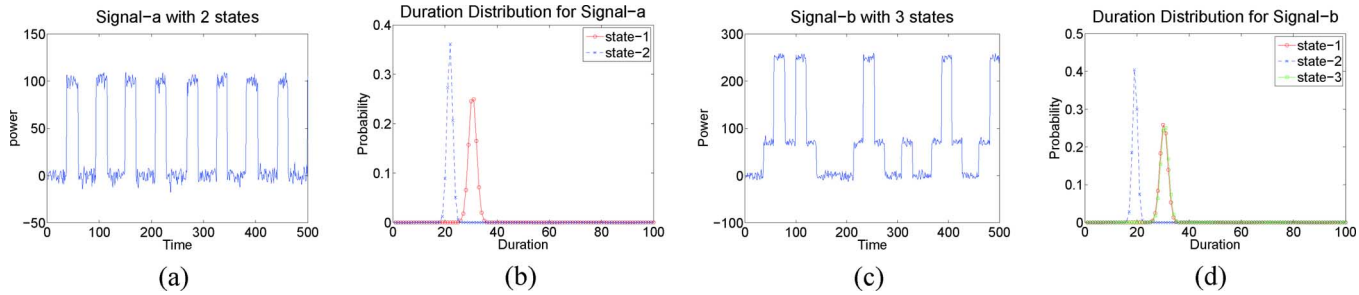


Fig. 4. We show the signals and the state duration distributions of the synthetic signals used the first set of experiments in Section III-A. Abstract indices are used for both axis without units. (a) Signal-a. (b) $P(d)$ for Signal-a. (c) Signal-b. (d) $P(d)$ for Signal-b.

TABLE II
ESTIMATION RESULTS OF HMM, EDHMM, AND THE PROPOSED EDHMM-DIFF ON THE SYNTHETIC DATA

Signal	Method	μ	$\hat{\mu}$	$d\mu$	$\hat{d\mu}$
Signal-a	**EDHMM-diff**	(0, 100.00)	(0, 100.37)	(30, 20)	(30, 21)
	HMM	(0, 100.00)	(108.18, 207.95)	(30, 20)	–
	EDHMM	(0, 100.00)	(107.96, 206.99)	(30, 20)	(31, 20)
Signal-b	**EDHMM-diff**	(0, 70, 250)	(0, 70.66, 249.67)	(30, 20, 30)	(28, 20, 32)
	HMM	(0, 70, 250)	(83.98, 153.94, 334.15)	(30, 20, 30)	–
	EDHMM	(0, 70, 250)	(83.99, 154.08, 330.57)	(30, 20, 30)	(31, 18, 30)

be consumed if there is a standby mode). We assume this assumption for all the experiments conducted in this paper.

We perform the estimation on Signal-a and Signal-b with the conventional HMM [4], EDHMM [10], and the proposed EDHMM-diff. In Table II, μ represents the true mean vector of the emission probability function that generates the signal, and $d\mu$ represents the true mean vector of the state durations. And $\hat{\mu}$ and $\hat{d\mu}$ are the estimated values, respectively. The estimated $\hat{P}(d)$'s for each state duration for Signal-a and Signal-b are plotted in Fig. 6. For the EDHMM-diff model, we initialize the model with $\mu = (0, 150)$ and $d\mu = (50, 50)$ for Signal-a; $\mu = (0, 100, 200)$ and $d\mu = (50, 50, 50)$ for Signal-b. From the results we can see that the proposed EDHMM-diff model can successfully estimate the power demands and state durations for multi-state power signals. It is worth pointing out that the estimated result shown in Fig. 6(b) seems different from the true distribution. Since a relatively large variance of 5 is added to the duration distribution of Signal-b to simulate the real-world situation, the proposed algorithm actually captures the true information in the signal, thanks to the non-parametric nature of the discrete Gaussian we used. The convergence of the proposed forward-backward algorithm is illustrated in Fig. 5 by showing the likelihood as a function of iteration index. It is clear that the proposed algorithm converges fast under both situations.

We also can see that the conventional HMM and EDHMM do a good job on estimating the "apparent" emission means. However, both of them cannot deal with the "aggregating effect", which is critical for estimating appliance loads from aggregated power signals. In this set of experiments, since the DC value added to Signal-a is generated randomly, the conventional HMM and EDHMM will give different (and incorrect) estimates for the same underlying load signals when different DC values are added.

In the second set of experiments, we demonstrate how to use the proposed EDHMM-diff model to perform detection, given

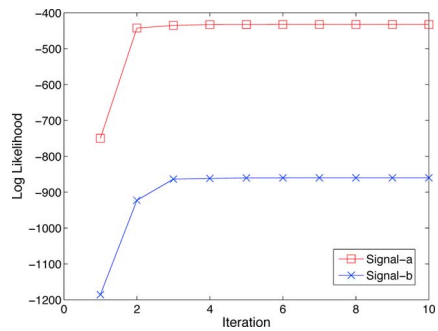


Fig. 5. Log-likelihood values as a function of the iteration index in the EDHMM-diff model when estimating Signal-a and Signal-b.

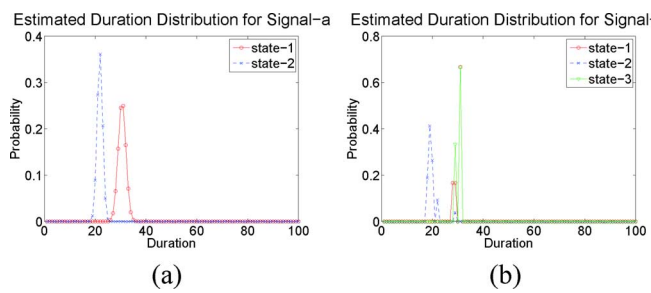


Fig. 6. (a) Estimated $\hat{P}(d)$ for Signal-a. (b) Estimated $\hat{P}(d)$ for Signal-b.

a specific appliance template. We would want our model to give higher log-likelihood scores for signal fragments which are similar to the template, and lower log-likelihood scores for signal fragments that are distinct from the template. It is worth noting that, random DC values are added to the generated signals to simulate the "aggregating effect". The template model $\mathcal{M}_0$ used here is with $\mu = (0, 150)$ and $d\mu = (50, 50)$, which is similar to Signal-a. The log-likelihood scores of the template given

TABLE III
LOG-LIKELIHOOD SCORES FOR DIFFERENT SIGNALS USED IN THE SECOND SET OF EXPERIMENTS IN SECTION III-A

	μ	$d\mu$	Log-likelihood
$\mathcal{M}_0$	(0, 150)	(50, 50)	–
Signal-a	(0, 100)	(30, 20)	**- 1.46**
Signal-a1	(0, 100)	(3, 2)	- 15.17
Signal-a2	(0, 100)	(50, 60)	**- 0.83**
Signal-a3	(0, 500)	(30, 20)	- 25.32
Signal-a4	(0, 200)	(30, 20)	**- 2.20**

different signals are shown in Table III, along with the parameters of the signals. From the table, we can see that Siganl-a, Signal-a3, and Signal-a4 get higher log-likelihood scores due to their similarity to the template model. We can also see that the EDHMM-diff model takes both power demands and state duration distributions into consideration. For example, Signal-a1 has similar μ but distinct $d\mu$, and Signal-a3 has similar $d\mu$ but distinct μ, while both get lower log-likelihood scores. Therefore, we can use the log-likelihood score estimated by the template model to detect similar signal fragments from aggregated power signals.

B. Experiments on Real Data

The main motivation of the proposed EDHMM-diff model is to estimate individual home appliance loads from the aggregated power signals. In this section, we test on aggregated real power signals collected from real houses. We use *refrigerator* and *dryer* as AOI examples in the following experiments, and the EDHMM-diff can be generalized to other appliances if needed. Both Reference Energy Disaggregation Data Set (REDD) and our own data collected by neurio™ system from the Energy Aware Technology Inc.[1] are investigated.

1) Experiments on Reference Energy Disaggregation Data Set (REDD): The REDD data set is proposed by [8], which contains both whole-home and circuit/device specific electric consumptions for a number of real houses over several months. To simulate the low frequency real power signals that we can usually access through smart meters, here we only use the aggregated whole-home real power signal and down-sample it to $(1/30)$ Hz (1 reading per 30 s) in our experiments. Although they claimed the breakdown signals were provided, those are actually circuit level signals and have corruptions for most of the devices, which results in no ground truth for our estimations.

As discussed in Section III-C, a template model is required for the detection purpose before we can extract the AOIs from aggregated signals. There are two ways to construct the templates: 1) estimate models from real signals by monitoring a number of AOI's and then average the models; 2) set the parameters in the template manually, according to the specifications reported by agencies such as Electric Power Research Institute (EPRI). We take the second approach in our experiments due to its simplicity. To demonstrate the robustness of the proposed method, we use the same template models for the same devices across all the houses. The template for **refrigerator** is $\mu = (0, 150), \sigma = (10, 10), d\mu = (50, 50), d\sigma = (5, 5), A = [0, 1; 1, 0], \pi = [1, 0]$. The template for **dryer** is $\mu(0, 1000), \sigma = (50, 100), d\mu = (1, 1), d\sigma = (5, 5), A = [0, 1; 1, 0], \pi = [1, 0]$.

[1]https://www.neur.io/

TABLE IV
ESTIMATION RESULTS ON THE REDD DATA SET. μ DENOTES THE GROUND TRUTH OF THE MEAN OF POWER DEMAND, $\hat{\mu}$ DENOTES THE ESTIMATED VALUE, AND $\hat{d\mu}$ DENOTES THE ESTIMATED MEAN OF DURATION DISTRIBUTION

House ID	Refrigerator		
	μ	$\hat{\mu}$	$\hat{d\mu}$
1	(0, 187.90)	(0, 180.56)	((17, 39, 57), 19)
2	(0, 243.26)	(0, 243.08)	((56, 64, 68, 100), 38)
3	(0, 122.23)	(0, 114.26)	((37, 50), (25, 28, 34))
4	(0, 113.67)	(0, 111.40)	((80, 84), 44)
5	(0, 137.29)	(0, 138.72)	((76, 95), 33)
6	(0, 153.26)	(0, 151.17)	((29, 35, 39, 45), 37)
	Cloth Dryer		
	μ	$\hat{\mu}$	$\hat{d\mu}$
1	(0, 1646.51)	(0, 1511.96)	((1, 10), 2)
3	(0, 2229.74)	(0, 2240.87)	(2, 3)

We set the length of the sliding window to be 50 min (100 points at the current granularity), and slide the windows for every 25 min (50 points correspondingly). After computing the likelihood scores of the template model for individual windows, we select the top windows with likelihood scores above a threshold as training signals for the re-estimation purpose.

We apply the "detect and re-estimate" procedure for *refrigerator* and *cloth dryer* on the signals from 6 houses in the REDD data set. The estimation results are reported in Table IV. It is worth noting that, since we model the state duration distribution as a discrete Gaussian function, for some cases, we obtain Gaussian mixtures for $P(d)$ with multiple centers. We report all these centers of mixtures in Table IV. The Gaussian mixtures actually give better estimations of the state durations than a single Gaussian density function, since some states of a certain appliance might have different durations. In addition, except house_1 and house_3, there are no valid cloth dryer signals in other houses.

As mentioned above, there is no reliable ground truth for the REDD data set, so that we manually calculated the means of power demands for all the devices as ground truth. Since it is hard to manually determine the duration distributions, we didn't report ground truth in the table. To verify the performance of the proposed method, we plot the detected signals of the refrigerators and cloth dryers from individual houses, along with the estimated state duration distributions in Fig. 8. By manually examining the signals, we can see that the estimation results listed in Table IV are reasonable for the home appliance load modeling purpose.

2) Experiments on the Energy Aware Data: The Energy Aware data is collected by the nerio™ system monitoring one apartment in Vancouver. There are around 10 appliances running in the apartment and several major appliances are sub-metered. To simulate the low frequency real power signals that we can usually access through smart meters, here we only use the aggregated whole-home real power signal and down-sample it to $(1/30)$ Hz (1 reading per 30 s) in our experiments. We follow the same experiment protocol as in the previous section, and perform "detection and re-estimation" for the refrigerator and the dryer as AOIs. For the refrigerator,

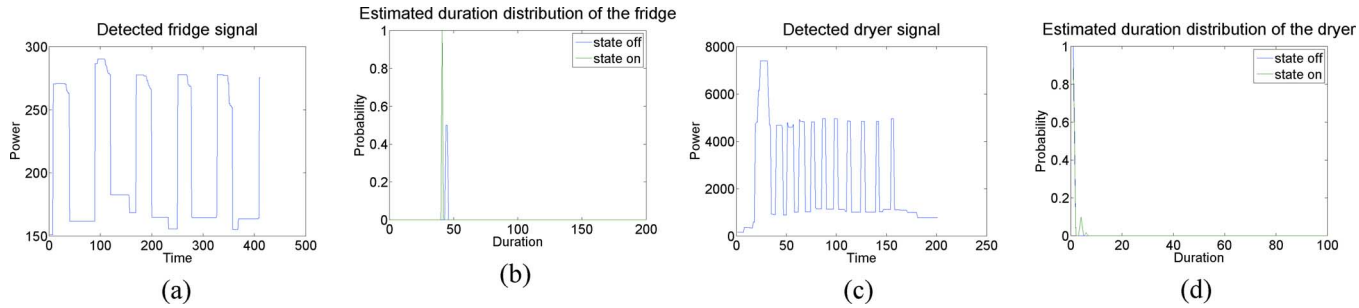


Fig. 7. Detected signals and corresponding estimated duration distributions for Energy Aware data. Y axis is power in Watts, and X axis is time in a unit of a half minute. (a) Refrigerator signal. (b) $P(d)$ for refrigerator signal. (c) Dryer signal. (d) $P(d)$ for dryer signal.

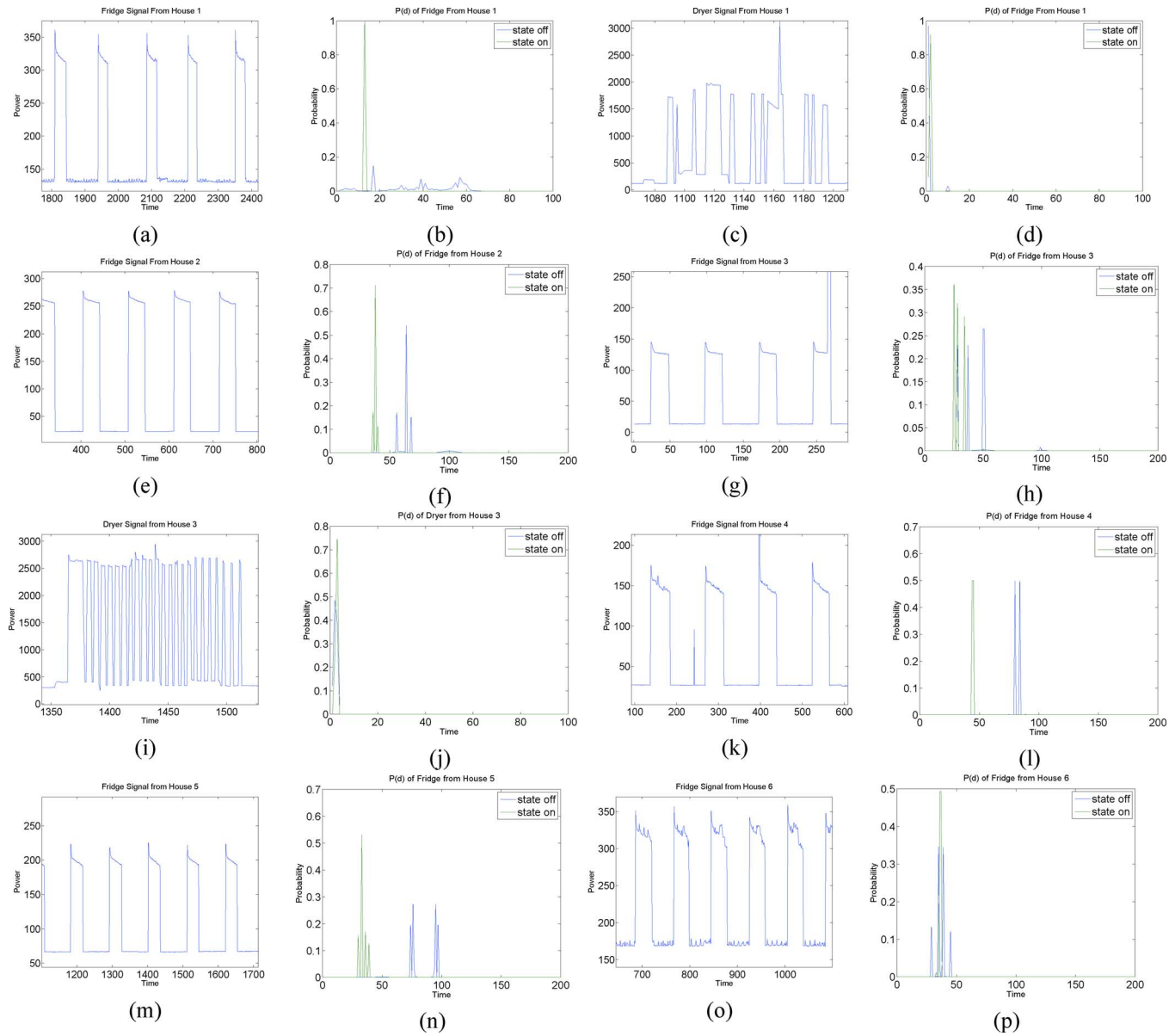


Fig. 8. REDD data set: Detected signal fragments and corresponding estimated state duration distributions for different appliances, where house_1 *refrigerator*: (a)(b), house_1 *dryer*: (c)(d), house_2 *refrigerator*: (e)(f), house_3 *refrigerator*: (g)(h), house_3 *dryer*: (i)(j), house_4 *refrigerator*: (k)(l), house_5 *refrigerator*: (m)(n), and house_6 *refrigerator*: (o)(p). Y axis is power in Watts, and X axis is time in a unit of a half minute.

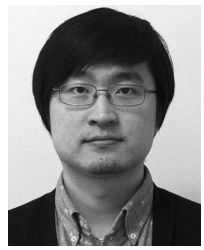
the estimated $\hat{\mu} = (0, 114.73)$, and for the dryer the estimated $\hat{\mu} = (0, 3786.7)$. The detected signals of the refrigerator and the dryer and their corresponding estimated duration distributions are shown in Fig. 7.

IV. Conclusion

In this paper, we tackle the appliance load modeling problem based on aggregated smart meter data by proposing an EDHMM-diff model and a specialized forward-backward inference and estimation algorithm. The proposed method can successfully model the state durations and overcome the problem caused by "aggregating effect". We demonstrate the effectiveness of the proposed method on synthetic data. The estimation results on real data, the REDD data set and Energy Aware data, show that the proposed EDHMM-diff model can be a promising solution for home appliance load modeling when only observing aggregated real power signals.

References

[1] G. W. Hart, "Nonintrusive appliance load monitoring," *Proc. IEEE*, vol. 80, no. 12, pp. 1870–1891, Dec. 1992.
[2] J. Liang, S. Ng, G. Kendall, and J. Cheng, "Load signature study Part I: Basic concept, structure, and methodology," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 551–560, Apr. 2010.
[3] J. Liang, S. K. Ng, G. Kendall, and J. W. Cheng, "Load signature study Part II: Disaggregation framework, simulation, and applications," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 561–569, Apr. 2010.
[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[5] J. Z. Kolter and T. Jaakkola, "Approximate inference in additive factorial HMMs with application to energy disaggregation," in *Proc. Int. Conf. Artificial Intelligence and Statistics*, 2012, pp. 1472–1482.
[6] O. Parson, S. Ghosh, M. Weal, and A. Rogers, "Non-intrusive load monitoring using prior models of general appliance types," in *Proc. 26th AAAI Conf. Artificial Intelligence*, 2012.
[7] H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han, "Unsupervised disaggregation of low frequency power measurements," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2010.
[8] J. Z. Kolter and M. J. Johnson, "REDD: A public data set for energy disaggregation research," in *Proc. SustKDD Workshop Data Mining Applications in Sustainability*, 2011, pp. 1–6.
[9] S.-Z. Yu and H. Kobayashi, "Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden Markov model," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1947–1951, May 2006.
[10] S.-Z. Yu, "Hidden semi-Markov models," *Artif. Intell.*, vol. 174, no. 4, pp. 215–243, 2010.



Zhenyu Guo received the B.E. degree from Zhejiang University, China, in 2009, and the Ph.D. degree from University of British Columbia, Vancouver, BC, Canada, in 2014.

He is a data scientist at Energy Aware Technology Inc., leading the data science research. His research interests are in machine learning and statistical signal processing with applications in computer vision, signal processing, and energy data analysis.

Z. Jane Wang (S'01–M'02–SM'12) received the B.Sc. degree from Tsinghua University, Beijing, China, in 1996, and the Ph.D. degree from the University of Connecticut, Storrs, CT, USA, in 2002.

She joined the Electrical and Computer Engineering Department, University of British Columbia, Vancouver, BC, Canada, in 2004, where she is an Associate Professor. Her research interests are in statistical signal processing theory and applications.

Dr. Wang is a corecipient of the EURASIP Best Paper Award 2004 and the IEEE Signal Processing Society Best Paper Award 2005. She has been an Associate Editor for several IEEE journals.

Ali Kashani received the Ph.D. degree in robotics from the University of British Columbia, Vancouver, BC, Canada, in 2011.

Since then, he has been with Energy Aware Technology, Vancouver, where he is currently leads the R&D.