A Bayesian Perspective on Residential Demand Response Using Smart Meter Data

Datong Zhou, Maximilian Balandat, and Claire Tomlin

Abstract— The widespread deployment of Advanced Metering Infrastructure has made granular data of residential electricity consumption available on a large scale. One field of research that relies on such granular consumption data is Residential Demand Response, where individual users are incentivized to temporarily reduce their consumption during periods of high marginal cost of electricity. To quantify the economic potential of Residential Demand Response, it is important to estimate the reductions during Demand Response hours, taking into account the heterogeneity of electricity users. In this paper, we incorporate latent variables representing behavioral archetypes of electricity users into the process of short-term load forecasting with Machine Learning methods, thereby differentiating between varying levels of energy consumption. The latent variables are constructed by fitting Conditional Mixture Models of Linear Regressions and Hidden Markov Models on smart meter data of a Residential Demand Response program in the western United States. We observe a notable increase in the accuracy of short-term load forecasts compared to the case without latent variables. We estimate the reductions during Demand Response events conditional on the latent variables and discover a higher DR reduction among users with automated smart home devices compared to those without.

I. INTRODUCTION

Residential Demand Response (DR) is a novel data-driven service enabled by the large-scale deployment of Advanced Metering Infrastructure (AMI). By communicating a proxy of the marginal price of electricity to consumers, it is acknowledged that economic efficiency can be increased [1]. During times when the grid is strained, a DR provider, which serves as a mediating unit between residential electricity consumers and the DR market, bids reductions with respect to an expected consumption (baseline) into the wholesale electricity market. Different market regulators, including CAISO, have launched such pilot programs [2], [3]. If the bid is cleared, the DR provider then prompts residential customers to temporarily reduce their consumption in exchange for a monetary reward proportional to the estimated reduction during DR times. As it is impossible to observe both the consumption conditional on DR-treatment and Non-DR-treatment, it becomes essential to estimate the counterfactual consumption, i.e. the consumption during DR times that would have been observed if no treatment had occurred. This is an application of the "Fundamental Problem of Causal Inference" [4], which states that it is impossible to observe more than one treatment on the same subject at one time.

For economic purposes, it is of cardinal importance for DR providers to bid the right amount of reductions into the wholesale electricity market, since penalties incur for shortfalls from the bidded capacity, and a suboptimal revenue would be recorded for a too modest bid. Assuming the bid is cleared, the major uncertainty is found to be the user behavior during DR times, i.e. the amount of reduction in response to the DR treatment. In [5], the authors find a positive correlation between the variability in consumption behavior and the magnitude of DR reduction, which suggests targeting variable households for a higher reduction yield.

In this paper, we analyze the heterogeneity in users' reduction behavior during DR times by using latent variables in statistical forecasting methods. This Bayesian perspective allows us to postulate the existence of behavioral archetypes of users, which govern the resulting and observable energy consumption. The latent variables are constructed in two ways: Firstly, we use a Conditional Gaussian Mixture Model (CGMM) of Linear Regressions, where the latent variable of a given data point is a vector of probabilities, with each component indicating the probability that the data point was generated by the corresponding mixture component. Secondly, we implement a Hidden Markov Model (HMM) whose hidden layer encodes hourly binary latent variables representing high and low levels of consumption, which in turn can be interpreted as an indicator for occupancy. The recommendation to DR providers is to prompt users only during hours of believed presence at home, thereby improving efficiency of targeting. Using this differentiation between different magnitudes of consumption, we observe a stark contrast in the estimated reduction between periods of high and low consumption.

In the extant literature, short-term load forecasting (STLF) has been extensively studied with different approaches and on different levels of aggregations of users, ranging from the individual level to city-wide predictions [6], [7]. Statistical time series models [8], [9], standard parametric regression models such as Ordinary Least Squares, Lasso- and Ridge-Regression [7], and non-parametric methods including k-Nearest Neighbors, Support Vector Regression [10], and Neural Networks [11] have been evaluated with respect to different metrics for accuracy. Widely explored Bayesian Methods for STLF are Gaussian Processes [12], Bayesian Neural Network approaches, e.g. for input selection problems [13], and Kalman-Filtering methods [14] with Hybrid Neural Network extensions [15]. HMMs for STLF have been applied

Datong Zhou is with the Department of Mechanical Engineering, University of California, Berkeley, USA. `datong.zhou@berkeley.edu`

Maximilian Balandat and Claire Tomlin are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA. `[balandat,tomlin]@eecs.berkeley.edu`

primarily for the purpose of occupancy detection [16] and Nonintrusive Load Monitoring [17]. [18] and [19] utilize occupancy information to increase the energy efficiency of building operation. To the best of our knowledge, CGMMs have not been investigated for STLF.

The contribution of this paper is two-fold: First, it aims to explore the potential for improvement in the prediction accuracy obtained by incorporating latent variable information from CGMMs and HMMs as an additional covariate into regression models. Second, it provides insights into the reduction behavior of users conditional on their latent states. Both aspects can help the DR provider make more informed bids into the wholesale market by targeting only the most susceptible users. The remainder of this paper is structured as follows: In Section II, we briefly outline classical Machine Learning (ML) methods used for STLF. Sections III and IV describe technical details of CGMMs and HMMs tailored to the specific needs of STLF, followed by Section V, which outlines the procedure of incorporating the estimated latent variables into STLF. Section VI presents a framework for estimating counterfactual consumption, which allows for the computation of the magnitude of the reduction of electricity consumption during DR hours. A case study on both semi-synthetic and observational data is presented in Section VII. Chapter VIII concludes the paper.

II. FORECASTING METHODS

The following well-established forecasting methods, which regress the outcomes Y on the covariates X, are used in the remainder of this paper:

- Ordinary Least Squares Regression (OLS)
- k Nearest Neighbors Regression (KNN)
- Support Vector Regression (SVR)
- Decision Tree Regression (DT)

Due to space limitations, we omit in-depth descriptions of the forecasting techniques. The reader is referred to the full version of this paper [5] and the references therein.

Notation: Let $Y \in \mathbb{R}^N$ denote a column vector of N scalar outcomes $\{y_1, \ldots, y_N\}$, e.g. in our case electricity consumption, and $X \in \mathbb{R}^{N \times d}$ the design matrix whose k-th row represents the covariates $x_k \in \mathbb{R}^d$ associated with outcome y_k. Let y and x denote a generic outcome and its associated covariate vector, respectively.

III. MIXTURE MODELS

In this section, we describe the fitting procedure of CG-MMs on data that combine multiple linear regression models to act as an ensemble learner. Given a set of covariate-outcome pairs (in our case y_i denotes energy consumption),

$$\mathcal{D} = \{(x_i, y_i) : i = 1, \ldots, N\}, \tag{1}$$

the idea is to model the probability distribution of any observation y with corresponding covariates x as the output of an ensemble of linear regressions

$$\mathbb{P}(y|x, \underbrace{\mathbf{w}, \sigma^2, \pi}_{=:\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y|w_k \cdot x, \sigma^2), \tag{2}$$

where $\pi = \{\pi_1, \ldots, \pi_K\}$ and $\mathbf{w} = \{w_1, \ldots, w_K\}$ denote K mixing proportions with $\sum_{i=1}^{K} \pi_k = 1$ and the regression coefficients for each learner, respectively. σ^2 signifies the noise variance, where, according to [20], we make the following

Assumption 1: σ^2 is equal across all mixture components $k = 1, \ldots, K$.

Assumption 1 can be relaxed by using mixture-specific noise covariances $\{\sigma_1^2, \ldots, \sigma_K^2\}$, in which case (5a)−(5d) need to be modified.

A. Parameter Estimation

Given the training data $\mathcal{D}$, the Expectation-Maximization Algorithm (EM-Algorithm) [21], [20] allows us to derive an iterative procedure to learn the parameters $\theta = \{\{\pi_k\}_{k=1}^{K}, \{w_k\}_{k=1}^{K}, \sigma^2\}$. We first define the expected complete log likelihood $\ell(\theta|\mathcal{D}_c)$, where

$$\mathcal{D}_c = \{(x_i, y_i, z_i) : i = 1, \ldots, N\} \tag{3}$$

denotes the fully observed dataset whose latent variables $\{z_1, \ldots, z_N\}$ are assumed to be known. The latent variable belonging to x_i is a vector $z_i = [z_{i1}, \ldots, z_{iK}]^\top$, where z_{ik} denotes the probability that x_i was generated by mixture component k. The complete log-likelihood is

$$\ell(\theta|\mathcal{D}_c) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \left(\pi_k \mathcal{N}(y_i|w_k \cdot x_i, \sigma^2) \right) \tag{4}$$

under the assumption of known z_{ik}. The EM-Algorithm alternates between the E-Step, whose task is to determine the expected value of the latent variables $z_{ik}, 1 \leq i \leq N, 1 \leq k \leq K$ with respect to the conditional probability distribution (2), and the M-Step, which updates the parameters θ with the results from the E-Step by taking the derivative of the expected value of (4) with respect to the desired parameters θ. This is carried out iteratively until some convergence criterion is reached, i.e. the incremental increase of the expected complete log likelihood (4) falls below a threshold. The update steps for one iteration are as follows:

$$\hat{z}_{ik} = \frac{\hat{\pi}_k \mathcal{N}(y_i|\hat{w}_k \cdot x_i, \hat{\sigma}^2)}{\sum_{j=1}^{K} \hat{\pi}_j \mathcal{N}(y_i|\hat{w}_j \cdot x_i, \hat{\sigma}^2)}, \tag{5a}$$

$$\hat{\pi}_k = \frac{1}{N} \sum_{i=1}^{N} \hat{z}_{ik}, \tag{5b}$$

$$\hat{w}_k = \left[X^\top D X \right]^{-1} X^\top D Y, \ D = \text{diag}(\hat{z}_{1k}, \ldots, \hat{z}_{Nk}), \tag{5c}$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{z}_{ik} (y_i - \hat{w}_k \cdot x_i)^2, \tag{5d}$$

where we have to incorporate the constraint $\sum_{k=1}^{K} \hat{\pi}_k = 1$ as a Lagrange Multiplier in the derivation.

B. Predicting New Data

To predict the outcome $\hat{y}$ of an out-of-sample data point x, we suggest a different approach than is employed by [20]: Instead of using the estimated mixing proportions $\{\hat{\pi}_k\}_{k=1}^{K}$

as the weights for a convex combination of the estimated regression coefficients $\{\hat{w}_k\}_{k=1}^K$, we choose the weights as the estimated latent variables $\{\hat{z}_{jk}\}_{k=1}^K$ of x's nearest neighbor x_j:

$$j = \arg \min_{1 \leq i \leq N} \|x_i - x\|_2 \tag{6a}$$

$$\hat{y} = \sum_{k=1}^K \hat{z}_{jk} \hat{w}_k \cdot x \tag{6b}$$

The rationale behind this approach is to exploit potential spatial separation in the set of training data, i.e. the fact that different regions of the covariate space are best fit by a specific learner. By locating the nearest neighbor of x, the same set of weights that proved to be most accurate for the training of the data points in the region around x are to be used for the prediction of $\hat{y}$.

IV. Hidden Markov Models

In this section, we briefly outline the training procedure of HMMs. Figure 1 shows the graphical model of a standard HMM with a hidden layer (transparent nodes), representing latent variables, and observations (shaded nodes).
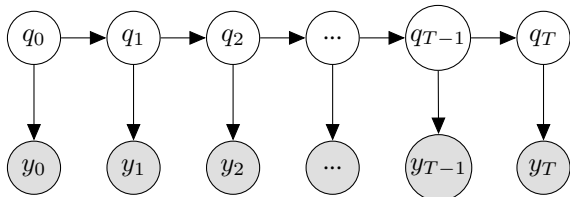

Fig. 1: Hidden Markov Model. Hidden States q, Observations y

A. Hidden Layer

We model the latent variables in the hidden layer (see Figure 1) as a first order, time-invariant, Discrete Time Markov Chain (DTMC) with a set of transition probabilities

$$a_{ij} = \mathbb{P}(q_t = j | q_{t-1} = i), \quad 1 \leq i, j \leq M, \tag{7}$$

where $t = 0, 1, 2, \ldots, T$ denote time instants associated with state changes and q_t the hidden state at time t. Due to the Markov Property, we have that, conditional on q_t, q_{t+1} is independent of q_{t-1}. The state transition coefficients a_{ij} have the properties

$$0 \leq a_{ij} \leq 1, \quad \sum_{j=1}^M a_{ij} = 1, \quad i, j \in \{1, \ldots, M\}, \tag{8}$$

where M denotes the number of states (=latent variables).

We postulate the existence of two different latent states for each hour of the day (HoD) between 6 a.m. - 8 p.m., and a single state for the remaining hours, hence $M = 38$. For the former hours, binary states describing each hour shall encode information about "high" ("H") or "low" ("L") consumption, which might be an indicator for occupancy ("H" = at home, "L" = not at home). For the remaining HoDs, we note that first, no DR events in our data set were recorded outside this window, and second, little variation in the smart meter recordings was observed, which is consistent with [5], where the authors find little variation in clustered load shapes during

the night. Due to the Markov Property, state transitions are restricted to states belonging to the next hour only, which renders the Markov transition matrix $A \in \mathbb{R}^{38 \times 38}$ sparse. Figure 2 shows the state transition diagram (without probabilities on the edges, which are to be estimated from data, see Section IV-C).
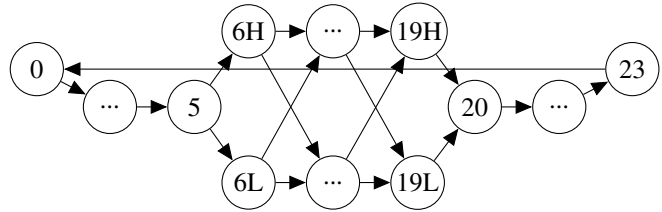

Fig. 2: Markov State Transition Diagram, 24 Hour Periodicity. For Example, "5" Signifies Time Between 5 a.m. - 6 a.m.

A logical extension is to allow for multi-step dependencies, which can be achieved by enlarging the state space of the DTMC such that the previous $n > 1$ states jointly determine the next transition. A more granular description of the state transitions, however, would come at the cost of a higher computational complexity, a tradeoff whose analysis is outside the scope of this paper.

A consequence of this modeling approach is that, if the consumption is high at time $t - 1$, it is likely that the hidden state $q_{t-1} = $ H and $q_t = $ H, and so we expect a high consumption at time t, as well. Conversely, if the consumption at time $t-1$ is low (i.e. due to an absent user), the most likely hidden state $q_{t-1} = $ L and $q_t = $ L, and thus we would expect a low consumption at time t. It turns out that the parameter estimation on the data set used in Section VII automatically assigns higher probabilities for transitions to the next hour of the same type than to the opposite type, indicating that switches between "H" and "L" do not occur frequently. This is consistent with our intuition: If the latent variable represents periods of expected presence or absence at home, users are more likely to remain either at home or absent, rather than switching every hour.

B. Observations

Assumption 2: Conditional on the current hidden state q_t, the observable energy consumption y_t (=observation/emission) is assumed to be normally distributed with parameters $(\mu_{q_t}, \sigma_{q_t}^2)$:

$$\mathbb{P}(y_t | q_t) = \frac{1}{\sqrt{2\pi\sigma_{q_t}^2}} \exp\left(-\frac{(y_t - \mu_{q_t})^2}{2\sigma_{q_t}^2}\right). \tag{9}$$

An obvious extension is to choose alternative distributions, an idea we do not investigate further in this paper.

C. Parameter Estimation and Inference

Given an observed sequence of emissions $Y := \{y_0, y_1, \ldots, y_T\}$ with known initial state distribution π_{q_0}, the parameters of the HMM $\theta := \{\{a_{ij}\}, \{\mu_{q_t}\}, \{\sigma_{q_t}^2\}\}$, i.e. the transition probabilities and emission parameters, can be estimated with the EM-Algorithm. Starting from the

complete log-likelihood

$$\ell(\theta|\mathcal{D}_c) = \log\left(\pi_{q_0}\prod_{t=0}^{T-1}a_{q_t,q_{t+1}}\prod_{t=0}^{T}\mathcal{N}(y_t|\mu_{q_t},\sigma_{q_t}^2)\right), \tag{10}$$

with the fully observed data set

$$\mathcal{D}_c = \{(y_n, q_n, a_{q_n,q_{n+1}}) : n \in [0, T-1]\} \cup \{\pi_{q_0}, y_T, q_T\}, \tag{11}$$

minimizing the expected value of (10) with respect to the desired variables θ to be estimated yields the update equations for the M-Step of the EM-algorithm (also called Baum-Welch Updates):

$$\hat{\pi}_i = \mathbb{P}(q_0 = i|Y) \tag{12a}$$

$$\hat{a}_{ij} = \frac{\sum_{t=0}^{T-1}\mathbb{P}(q_t = i, q_{t+1} = j|Y)}{\sum_{t=0}^{T-1}\sum_{j=1}^{M}\mathbb{P}(q_t = i, q_{t+1} = j|Y)} \tag{12b}$$

$$\hat{\mu}_i = \frac{\sum_{t=0}^{T} y_t \cdot \mathbb{P}(q_t = i|Y)}{\sum_{t=0}^{T}\mathbb{P}(q_t = i|Y)} \tag{12c}$$

$$\hat{\sigma}_i^2 = \frac{\sum_{t=0}^{T}\mathbb{P}(q_t = i|Y)(y_t - \hat{\mu}_i)^2}{\sum_{t=0}^{T}\mathbb{P}(q_t = i|Y)} \tag{12d}$$

To arrive at Equations (12a) and (12b), the stochastic constraints described in (8) and sparsity patterns of the transition matrix A as well as $\sum_{i=1}^{M}\pi_i = 1$ are used as Lagrange multipliers during the minimization of (10).

Using Bayes Rule, the E-Step of the EM-algorithm computes the sufficient statistics $\mathbb{P}(q_t = i, q_{t+1} = j|Y)$ and $\mathbb{P}(q_t = i|Y)$ with the well-known *Alpha-Beta-Recursion*:

$$\mathbb{P}(q_t|Y) = \frac{\mathbb{P}(Y|q_t)\mathbb{P}(q_t)}{\mathbb{P}(Y)}$$
$$= \frac{\mathbb{P}(y_0,\ldots,y_{t-1},q_t)\mathbb{P}(y_t|q_t)\mathbb{P}(y_{t+1},\ldots,y_T|q_t)}{\mathbb{P}(Y)}$$
$$=: \frac{\alpha(q_t)\mathbb{P}(y_t|q_t)\beta(q_t)}{\mathbb{P}(Y)}. \tag{13}$$

We note that $\alpha(q_t)$ is defined as $\mathbb{P}(y_0,\ldots,y_{t-1},q_t)$ rather than $\mathbb{P}(y_0,\ldots,y_t,q_t)$ as is done in [21], [22]. This is done for a simplified treatment of its update step (14) and the prediction problem (18).

Using Bayes Rule, $\alpha(q_t)$ and $\beta(q_t)$ can be updated recursively:

$$\alpha(q_{t+1}) = \mathbb{P}(y_0,\ldots,y_t,q_{t+1})$$
$$= \sum_{q_t}\mathbb{P}(y_0,\ldots,y_t,q_t,q_{t+1})$$
$$= \sum_{q_t}\alpha(q_t)\mathbb{P}(y_t|q_t)a_{q_t,q_{t+1}}. \tag{14}$$

$$\beta(q_t) = \mathbb{P}(y_{t+1},\ldots,y_T|q_t)$$
$$= \sum_{q_{t+1}}\mathbb{P}(y_{t+1},\ldots,y_T,q_{t+1}|q_t)$$
$$= \sum_{q_{t+1}}\beta(q_{t+1})\mathbb{P}(y_{t+1}|q_{t+1})a_{q_t,q_{t+1}}. \tag{15}$$

Note that $\mathbb{P}(y_t|q_t),\ 0 \leq t \leq T$ can be computed with (9). $\alpha(q_1)$ is initialized as π_{q_0} and $\beta(q_T)$ as a vector of ones.

With the definition of $\alpha(q_t)$ and $\beta(q_t)$, $\mathbb{P}(q_t, q_{t+1}|Y)$ is computed as follows:

$$\mathbb{P}(q_t, q_{t+1}|Y) = \frac{\mathbb{P}(Y|q_t,q_{t+1})\mathbb{P}(q_t,q_{t+1})}{\mathbb{P}(Y)}$$
$$= \frac{\alpha(q_t)\beta(q_{t+1})a_{q_t,q_{t+1}}\mathbb{P}(y_t|q_t)\mathbb{P}(y_{t+1}|q_{t+1})}{\mathbb{P}(Y)}. \tag{16}$$

In summary, the EM-algorithm iterates between the E-Step to compute the sufficient statistics $\mathbb{P}(q_t = i, q_{t+1} = j|Y)$ and $\mathbb{P}(q_t = i|Y)$ with Equations (13), (14), (15), and (16) while fixing the parameters in (12a)−(12d), and the M-Step to update the parameters in (12a)−(12d) while fixing the sufficient statistics until some convergence criterion on the expected value of (10) is reached.

D. Filtering, Smoothing, and Predicting the Latent Variable

After the parameters of the HMM have been estimated, we turn to the problem of estimating the probabilities of the most likely hidden state. Given the observation sequence $Y := \{y_0, y_1, \ldots, y_T\}$, the *filtering problem* calculates $\mathbb{P}(q_T|Y)$:

$$\mathbb{P}(q_T|y_0,\ldots,y_T) = \frac{\mathbb{P}(y_0,\ldots,y_T|q_T)\mathbb{P}(q_T)}{\mathbb{P}(y_0,\ldots,y_T)}$$
$$= \frac{\alpha(q_T)\mathbb{P}(y_T|q_T)}{\mathbb{P}(y_0,\ldots,y_T)}. \tag{17}$$

Alternatively, the *prediction problem* can be used to predict the probability of the next hidden state at time $T+1$, i.e.

$$\mathbb{P}(q_{T+1}|y_0,\ldots,y_T) = \frac{\mathbb{P}(y_0,\ldots,y_T|q_{T+1})\mathbb{P}(q_{T+1})}{\mathbb{P}(y_0,\ldots,y_T)}$$
$$= \frac{\alpha(q_{T+1})}{\mathbb{P}(y_0,\ldots,y_T)}. \tag{18}$$

Lastly, the *smoothing problem* can be solved to ex-post predict the probability of the latent variable at a past time $1 \leq p < T$:

$$\mathbb{P}(q_p|y_0,\ldots,y_T) = \frac{\mathbb{P}(y_0,\ldots,y_T|q_p)\mathbb{P}(q_p)}{\mathbb{P}(y_0,\ldots,y_T)}$$
$$= \frac{\alpha(q_p)\mathbb{P}(y_p|q_p)\beta(q_p)}{\mathbb{P}(y_0,\ldots,y_T)}. \tag{19}$$

V. SHORT-TERM LOAD FORECASTING

In the following, we describe *online* forecasting algorithms that allow for including knowledge about the estimated latent variables obtained from HMMs and CGMMs into the ML methods introduced in Section II. We make the following

Assumption 3: The consumption time series Y is stationary, i.e. there are no structural changes in consumption behavior over time.

This assumption is sound as we explain in Section VII.

A. Covariates for Prediction

The following observable covariates are used for all forecasting methods:

- Five previous hourly consumptions
- Five previous hourly ambient air temperatures
- A categorical variable for the hour of day for ML methods without latent variable and the CGMM

- A categorical variable interacting the hour of day with the estimated latent variable obtained from HMM for ML methods with HMM

B. Prediction with Hidden Markov Model

Algorithm 1 Algorithm for Online Prediction with HMM

Input: Training Data $\mathcal{D}_{\mathrm{tr}} := \{(x_t, y_t) : t = 0, \ldots, T\}$, Test Data $\mathcal{D}_{\mathrm{te}} := \{(x_t, y_t) : t = T + 1, \ldots, \tau\}$, ML Method

1: Initialize all $\mu_1, \ldots, \mu_{38}, \sigma_1^2, \ldots, \sigma_{38}^2$ suitably
2: Initialize all a_{ij}, observing (8) and Figure 2
3: **while** $\Delta \mathbb{E}\left[\ell(\theta|\mathcal{D}_c)\right] < \epsilon$ **do**
4:     Do E-Step: Calculate (9) and (16) for $t = [0, \ldots, T-1]$, $q_t, q_{t+1} = [1, \ldots, 38]$ with (13)−(15)
5:     Do M-Step: Update HMM parameters with (12a)−(12d)
6: **end while**
7: Solve smoothing problem (19) for $t = 0, \ldots, T-1$
8: Solve filtering problem (17) for $t = T$
9: Round $\mathbb{P}(\hat{q}_0|\mathcal{D}_{\mathrm{tr}}), \ldots, \mathbb{P}(\hat{q}_T|\mathcal{D}_{\mathrm{tr}})$ to 0 / 1
10: Fit ML Method on $\{((x_t, \mathbb{P}(\hat{q}_t)), y_t) : t \in 0, \ldots, T\}$
11: **for** s in $[T+1, \tau]$ **do**
12:     Solve prediction problem (18) at time s
13:     Round $\mathbb{P}(\hat{q}_s)$ to 0 / 1
14:     Predict $\hat{y}_s$ with ML method on covariates $(x_s, \mathbb{P}(\hat{q}_s))$
15: **end for**
16: **return** $\hat{y}_{T+1}, \ldots, \hat{y}_\tau$

Algorithm 1 describes the procedure of fitting an HMM on training data $\mathcal{D}_{\mathrm{tr}}$, which yields estimated latent variables to be used as additional covariates for the ML methods presented in Section II to perform stepwise prediction on the covariates of the test data $\mathcal{D}_{\mathrm{te}}$. The prediction accuracy of these outcomes is then compared to those outcomes predicted by ML methods that are trained on the training data $\mathcal{D}_{\mathrm{tr}}$ without estimated latent variables in the covariates.

C. Prediction with Conditional Gaussian Mixture Model

Algorithm 2 describes the online prediction method for a CGMM with $k = 2$ on a given set of training and test data. $\hat{w}$ obtained by OLS is perturbed with zero mean Gaussian Noise ϵ to obtain the initializations w_1, w_2. Note that this step is necessary to break the symmetry of the update steps (5a)−(5d), which would keep $w_1 = w_2 = \hat{w}$ unchanged.

Note that in both Algorithms 1 and 2, the model-specific parameters could be updated after each prediction as more data from the test sequence is observed and hence enters $\mathcal{D}_{\mathrm{tr}}$.

D. Metric for Forecasting Accuracy

The Mean Absolute Percentage Error (MAPE) of predictions of a set of discrete values $v_i \in \mathcal{V}$ is used to evaluate the accuracy of the predictor:

$$\mathrm{MAPE} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left| \frac{\hat{v}_i - v_i}{v_i} \right| \cdot 100\%, \qquad (20)$$

where $\hat{v}_i$ denotes the estimate of v_i.

Algorithm 2 Algorithm for Online Prediction with CGMM

Input: Training Data $\mathcal{D}_{\mathrm{tr}} := \{(x_t, y_t) : t = 0, \ldots, T\}$, Test Data $\mathcal{D}_{\mathrm{te}} := \{(x_t, y_t) : t = T + 1, \ldots, \tau\}$

1: Fit OLS model on $\mathcal{D}_{\mathrm{tr}}$ to obtain $\hat{w}$
2: Initialize $w_1 \leftarrow \hat{w} + \epsilon$
3: Initialize $w_2 \leftarrow \hat{w} + \epsilon$
4: **while** $\Delta \mathbb{E}\left[\ell(\theta|\mathcal{D}_c)\right] < \epsilon$ **do**
5:     Update CGMM parameters (5a)−(5d)
6: **end while**
7: **for** s in $[T+1, \tau]$ **do**
8:     Predict $\hat{y}_s$ with (6a) and (6b)
9: **end for**
10: **return** $\hat{y}_{T+1}, \ldots, \hat{y}_\tau$

VI. NON-EXPERIMENTAL ESTIMATES OF DR REDUCTION

To estimate individual treatment effects, we adopt the potential outcomes framework [23] with binary treatments $T_t \in \{0, 1\}$, where $T_t = 1$ corresponds to a DR intervention at time t, and $T_t = 0$ denotes its absence, hence "control". Let y_t^0 and y_t^1 denote the response (i.e. the electricity consumption) that would be observed if an individual received treatment 0 and 1 at time t, respectively. The goal is to estimate the conditional treatment effect, i.e.

$$\Delta(x) = \mathbb{E}\left[y^1 | x \in \mathcal{X}\right] - \mathbb{E}\left[y^0 | x \in \mathcal{X}\right], \qquad (21)$$

where x denotes a vector of observable covariates in the covariate space $\mathcal{X}$. Assuming an unconfounded assignment mechanism of treatments to individuals and independency of the potential outcomes of time, then, conditional on the covariates (see [23] for details), the true causal effect of DR, namely $\left(y_t^0 - y_t^1\right)$, cannot be found because only one of y_t^0 and y_t^1 can be observed (c.f. Fundamental Problem of Causal Inference [4]).

Thus, Causal Inference is a "Missing Data Problem". Given the observed treatment outcomes $y_{t_1}^1, \ldots, y_{t_n}^1$, to estimate the true causal effect of treatment, one would require a credible estimate of the *counterfactuals* $\hat{y}_{t_1}^0, \ldots, \hat{y}_{t_n}^0$, i.e. the outcomes in the hypothetical absence of treatment.

To compute such estimates in a non-experimental way, we split the available consumption data into a pretreatment training set with time indices $t \in \mathcal{P}$ consisting of "regular" electricity consumption, i.e. all smart meter readings before the customers' signup date with the DR provider, and a test set with corresponding times $t \in \mathcal{S}$ thereafter which itself consists of smart meter readings during DR hours $\mathcal{T}$ (treatment) and outside DR hours $\mathcal{C}$ (control), hence $\mathcal{S} = \mathcal{T} \cup \mathcal{C}$. Let

$$\mathcal{D}_P = \{(x_{i,t}^0, y_{i,t}^0) : t \in \mathcal{P}\} \qquad (22a)$$
$$\mathcal{D}_C = \{(x_{i,t}^0, y_{i,t}^0) : t \in \mathcal{C}\} \qquad (22b)$$
$$\mathcal{D}_T = \{(x_{i,t}^1, y_{i,t}^1) : t \in \mathcal{T}\} \qquad (22c)$$

denote covariate/outcome pairs for the pretreatment period, the control observations, and the treatment observations of

user i, respectively. By fitting any regression model presented in Section II on the pretreatment training data $\mathcal{D}_P$ of a given user i, and under Assumption 3, applying this model on the treatment covariates $\{x_{i,t}^1 : t \in \mathcal{T}\}$ yields user i's estimated counterfactuals $\{\hat{y}_{i,t}^0 : t \in \mathcal{T}\}$. In particular, Assumption 3 states that DR treatments are interpreted as transitory shocks that do not result in a change in the consumption behavior for $t \in \mathcal{C}$. An elementwise comparison of $\{\hat{y}_{i,t}^0 : t \in \mathcal{T}\}$ and $\{y_{i,t}^1 : t \in \mathcal{T}\}$ yields the pointwise estimated reduction of user i's consumption $\{\hat{y}_{i,t}^\Delta : t \in \mathcal{T}\}$ during each DR event:

$$\{\hat{y}_{i,t}^\Delta : t \in \mathcal{T}\} = \{(\hat{y}_{i,t}^0 - y_{i,t}^1) : t \in \mathcal{T}\}. \tag{23}$$

$\hat{y}_{i,t}^\Delta > 0$ corresponds to an estimated reduction of $\hat{y}_{i,t}^\Delta$, and conversely, $\hat{y}_{i,t}^\Delta < 0$ implies an estimated increase by $|\hat{y}_{i,t}^\Delta|$.

VII. EXPERIMENTS ON DATA

We conduct a case study on a data set of a residential DR program including residential customers in the western United States, collected between 2012 and 2014. Aligned with those readings are timestamps of notifications sent by the DR provider to the users that prompt them to reduce their consumption for a short period, typically until the next full hour. A subset of the users have smart home devices that can be remotely shut off by the DR provider with the users' consent. Ambient air temperature measurements were logged from publicly available data sources to capture the correlation between temperature and electricity consumption.

A. Characteristics of Data and Data Preprocessing

Users with the following characteristics are excluded from the analysis:
- Users with residential solar photovoltaics (PV)
- Users with corrupt smart meter readings, i.e. unrealistically high recordings

The consumption series of the remaining users are then aligned with available temperature readings and mapped to the range $[0,1]$ to be able to compare users on a relative level. The temperature data is standardized to zero mean and unit variance. Lastly, the pretreatment data is tested for stationarity with the augmented Dickey-Fuller Test [24] to assert, with a significance level of more than 99%, the absence of a unit root, which motivates Assumption 3.

B. Experiments on Semi-Synthetic Data

As only one of $\{y_{i,t}^0, y_{i,t}^1\}$ for a given user i at time t can be observed, we construct semisynthetic data for which both values and hence the true causal effect $(y_{i,t}^0 - y_{i,t}^1)$ are known. This allows us to evaluate the accuracy of predicted counterfactual consumptions and the ensuing non-experimental estimates of DR reduction (23). For this purpose, we take actual pretreatment training data $\mathcal{D}_P$ (22a) for each user i, which is free of any DR messages. Next, we split this training set into two pieces by introducing an artificial signup date $\tilde{t}$ valid across all users. We thus obtain a synthetic training set $\tilde{\mathcal{D}}_P = \{(x_{i,t}^0, y_{i,t}^0) : t \in \mathcal{P}, t < \tilde{t}\}$ and a synthetic test set $\tilde{\mathcal{D}}_S = \{(x_{i,t}^0, y_{i,t}^0) : t \in \mathcal{P}, t \geq \tilde{t}\}$ for user i. Next, a random subset $\tilde{\mathcal{T}}$ of all available time indices in the synthetic

test set $\tilde{\mathcal{D}}_S$ between 6 a.m. - 8 p.m. is assigned a synthetic treatment, for which the respective consumption is decreased by a random value $\in [0, \bar{c}]$. By doing so, both the treatment and control outcomes for $t \in \tilde{\mathcal{T}}$ become available, and so we obtain the semisynthetic data set

$$\tilde{\mathcal{D}}_{\tilde{T}} := \{(x_{i,t}^0, y_{i,t}^0, y_{i,t}^1) : t \in \tilde{\mathcal{T}}\}. \tag{24}$$

Thus, any non-experimental estimate of the DR treatment effect for $t \in \tilde{\mathcal{T}}$ can be benchmarked on the known (synthetic) counterfactual $\{y_{i,t}^0 : t \in \tilde{\mathcal{T}}\}$.

This semisynthetic data set is used for two purposes. First, we evaluate the MAPE (20) of the estimators from Section II, with and without latent variables. This is done by training them on $\mathcal{D}_{\text{tr}} = \tilde{D}_P$ and testing on $\mathcal{D}_{\text{te}} = \tilde{D}_S$, which yields out-of-sample counterfactual consumptions $\{\hat{y}_{i,t}^0 : t \in \tilde{\mathcal{T}}\}$ across all users i, see Algorithms 1 and 2. Second, we conduct a comparison of the eventwise errors of estimated DR reductions for all ML methods with the HMM latent variable (CGMM is not considered further), which, for a given user i at time t, are obtained as follows:

$$\hat{y}_{i,t}^\Delta - y_{i,t}^\Delta = (\hat{y}_{i,t}^0 - y_{i,t}^1) - (y_{i,t}^0 - y_{i,t}^1) = \hat{y}_{i,t}^0 - y_{i,t}^0. \tag{25}$$

The ground truth counterfactual $y_{i,t}^0$ is available for the semisynthetic data (24) by construction, but would be unavailable for real-world data.

Figure 3 shows a boxplot of the distribution of average MAPEs across users for the prediction methods introduced in Section II with and without the latent variable from HMM, and for the CGMM (Section III). It can be seen that the
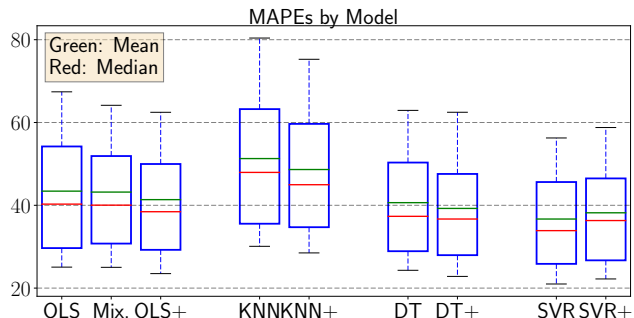
Fig. 3: Prediction Accuracy by Forecasting Method. "+" Signifies Model with HMM Latent Variable, "Mix." Denotes CGMM. Blue Boxes Span 25-75th Percentile, Whiskers 10-90th.

information about the latent variable improves the prediction accuracy in all cases but SVR. Further, the lower MAPE obtained with DT and SVR is consistent with the findings in [7], [9]. The higher MAPE for KNN compared to OLS can be explained by the different magnitudes of the covariates introduced in Section V-A, which gives categorical variables disproportionate weight. The CGMM performs better than OLS, but worse than OLS with the latent variable. Note that other more sophisticated predictors (e.g. Neural Networks) have lower MAPEs at the cost of longer computation times and potential loss of interpretability, but are likely to show a similar improvement in terms of MAPE by incorporating information about the estimated latent variable as the amount of training data increases. For a comparison between the

prediction accuracy of state-of-the-art estimators, the reader is referred to [7], [9] for further information.

Figure 4 shows histograms of eventwise prediction errors (25) in the estimated DR reduction for single events and across all users i. Green bars and red bars signify prediction errors of forecasting methods that do and do not make use of the estimated latent variable from HMM, respectively. Aligned with these plots are the sample mean and covariance of the errors for the models that take the latent variable into account. The bias-variance decomposition

$$\mathbb{E}\left[\left(\hat{y}_{i,t}^{\Delta} - y_{i,t}^{\Delta}\right)^2\right] = \mathrm{Bias}(\hat{y}_{i,t}^{\Delta})^2 + \mathrm{Var}(\hat{y}_{i,t}^{\Delta}) + \epsilon, \quad (26)$$

where ϵ denotes the irreducible error, is invoked in the following. Noting that $\hat{\mu}$ and $\hat{\sigma}^2$ in Figure 4 correspond to the bias and variance in (26) from the model with latent variable from HMM, the tradeoff becomes clear when comparing OLS, DT, and SVR. A lower variance of DT and SVR comes at the cost of a higher bias. For KNN, both bias and variance are larger than in OLS, which is explained by the poor predictive performance of KNN (see Figure 3). For a subsequent analysis of individual treatment effects (ITEs), we choose the least biased estimator that uses latent variables, in our case OLS, despite its higher overall prediction error compared to SVR and DT.
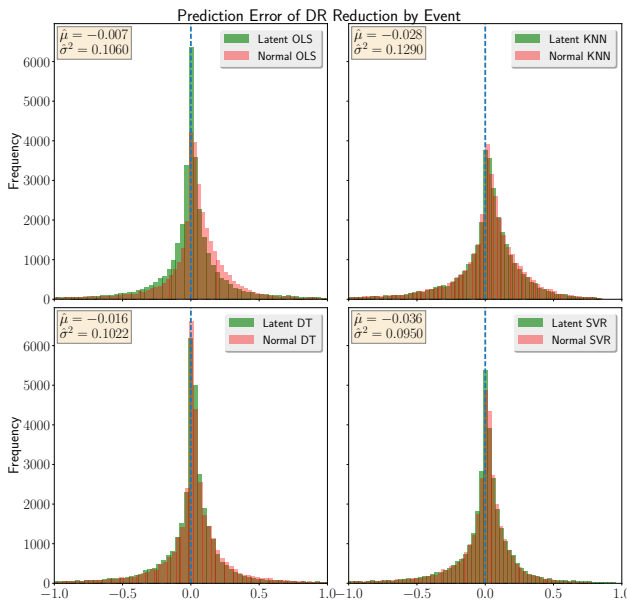

Fig. 4: Pointwise Prediction Error of DR Treatment Effect on the User Level; Bias $\hat{\mu}$ and Variance $\hat{\sigma}^2$ of Model with Latent Variable from HMM.

C. Experiments on Actual Data

In the following, we analyze ITEs for users with and without smart home devices. The analysis of reduction is carried out with OLS that utilizes an estimate of the HMM latent state because it is found that this method has the lowest bias on semisynthetic data (see Figure 4).

In the real data case, only the treatment outcomes $\{y_{i,t}^1 : t \in \mathcal{T}\}$ for user i are observed during DR events, and so the counterfactuals $\{\hat{y}_{i,t}^0 : t \in \mathcal{T}\}$ are predicted to calculate a non-experimental estimate of the DR reduction (23). Using

Algorithm 1 on the pre-signup data $\mathcal{D}_P$ (22a) as training data $\mathcal{D}_{\mathrm{tr}}$ for each user and $\mathcal{D}_{\mathrm{te}} = \mathcal{D}_C \cup \mathcal{D}_T$ (22b), (22c), the pointwise reductions across all users and each treatment $t \in \mathcal{T}$ are calculated. Figure 5 shows boxplots of estimated DR reductions conditional on (a) the hour of day, (b) users with and without smart home devices, and (c) the predicted latent states. The gray bars represent "placebo" events (i.e. a subset of hours $t \in \mathcal{C}$ outside DR treatments hours, but after the signup date) estimated by the same model.

Figure 5 gives rise to two observations: First, the estimated reduction conditional on the "high" latent state is greater in magnitude for users with smart home devices, following the intuition that the "high" state describes the operation of smart home devices which can be conveniently shut off during DR hours. In contrast, the lower estimated reductions of regular users during "high" latent states might reflect the additional hassle cost that incurs for users to manually reduce their consumption. Second, the estimated reductions for both users with and without smart home devices and conditional on the "low" latent state show mean reductions around zero, contrary to the expectation of a small positive reduction. This might indicate the existence of a threshold representing the standby consumption of users, below which it is hard or impossible to reduce consumption further.

This finding could be particularly meaningful to the DR provider, as it presents a recommendation as to when to call DR events and for which users, which could improve the allocative efficiency of DR targeting and be a stepping stone towards calculating optimal bids.

VIII. Conclusion

We developed non-experimental estimators from Machine Learning for estimating ITEs of Residential Demand Response and showed that incorporating a latent variable, either with a Conditional Gaussian Mixture Model or a Hidden Markov Model, allows for an improvement in prediction accuracy. This Bayesian approach is motivated by the need to obtain interpretable and physically meaningful results capturing the users' electricity consumption behavior. We then tested the forecasting algorithms on semi-synthetic data to find that Ordinary Least Squares in conjunction with a latent variable produces the least biased estimator for DR reduction. Lastly, this estimator was applied on a residential DR data set to determine hourly reductions of electricity consumption for both users with and without automated electric devices. The highest reductions were found to be among users with home automation devices during "high" estimated latent states, which in turn provides a recommendation for DR providers for targeting purposes, i.e. to focus on automated users for the highest yield in reduction.

This paper provides only a foundation for more profound analyses in the area of Residential Demand Response. In particular, latent variables can be added as an additional covariate to more computationally demanding estimators, for instance Neural Networks or Random Forests, in order to assess the gain in forecasting precision with latent variables. This is an area to be explored by the established area of
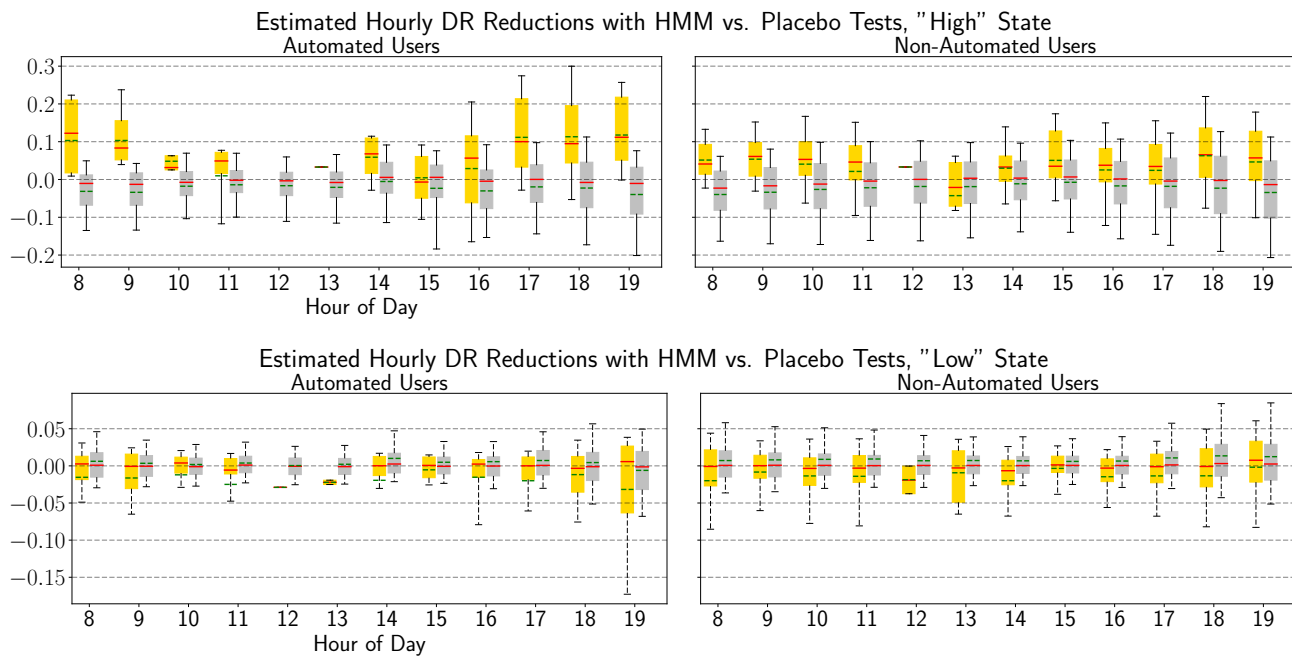
Fig. 5: Estimated Reduction Across Users by Hour of Day (Yellow) vs. Estimated Reduction for Placebo Events (Gray) for Automated and Non-Automated Users Conditional on Estimated Latent Variable. Red: Median, Green: Mean. Blue Boxes Span 25-75th Percentile, Whiskers 10-90th.

STLF, which has traditionally been focusing on maximizing the precision of forecasting algorithms. Further, various extensions to modeling the HMM are worth exploring, such as enlarging the state space of the Markov Chain to enforce a dependency on more than just the previous hour, or increasing the number of hidden states for a given hour (i.e "low", "medium", and "high" consumption). Lastly, the estimated latent variable could be related to a measure of occupancy in residential dwellings, and so a validation of the estimated latent states on ground truth data on occupancy would be interesting if privacy concerns could be overcome.

REFERENCES

[1] Federal Energy Regulatory Commission (FERC), "National Action Plan on Demand Response," June 2010.
[2] PJM Interconnection LLC, "PJM Capacity Performance Updated Proposal," Oct 2014.
[3] Public Utilities Commission of the State of California (CPUC), "Resolution E-4728. Approval with Modifications to the Joint Utility Proposal for a DR Auction Mechanism Pilot," July 2015.
[4] P. W. Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.
[5] D. Zhou, M. Balandat, and C. Tomlin, "Residential Demand Response Targeting Using Machine Learning with Observational Data," *55th Conference on Decision and Control, Available Online*, 2016.
[6] R. A. Sevlian and R. Rajagopal, "A Model For The Effect of Aggregation on Short Term Load Forecasting," *IEEE Transactions on Power Systems*, 2014.
[7] P. Mirowski, S. Chen, T. K. Ho, and C.-N. Yu, "Demand Forecasting in Smart Grids," *Bell Labs Technical Journal*, vol. 18, no. 4, 2014.
[8] S. Arora and J. W. Taylor, "Forecasting Electricity Smart Meter Data Using Conditional Kernel Density Estimation," *Omega*, 2014.
[9] J. W. Taylor and P. E. McSharry, "Short-Term Load Forecasting Methods: An Evaluation Based on European Data," *IEEE Transactions on Power Systems*, vol. 22, no. 4, 2007.
[10] E. E. Elattar, J. Goulermas, and Q. H. Wu, "Electric Load Forecasting Based on Locally Weighted Support Vector Regression," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 4, 2010.
[11] R. E. Edwards, J. New, and L. E. Parker, "Predicting Future Hourly Residential Electrical Consumption: A Machine Learning Case Study," *Energy and Buildings*, vol. 49, pp. 591–603, 2012.
[12] P. Lauret, M. David, and D. Caloigne, "Nonlinear Methods for STLF," *Energy Procedia*, vol. 14, pp. 1404–1409, 2012.
[13] H. S. Hippert and J. W. Taylor, "An Evaluation of Bayesian Techniques for Controlling Model Complexity and Selecting Inputs in a Neural Network for Short-Term Load Forecasting," *Neural Networks*, vol. 23, pp. 386–395, 2010.
[14] H. M. Al-Hamadi and S. A. Soliman, "Short-Term Electric Load Forecasting Based on Kalman Filtering Algorithm with Moving Window Weather and Load Model," *Electric Power Systems Research*, vol. 68, no. 1, pp. 47–59, 2004.
[15] C. Guan, P. B. Luh, L. D. Michel, and Z. Chi, "Hybrid Kalman Filters for Very STLF and Prediction Interval Estimation," *IEEE Transactions on Power Systems*, vol. 28, no. 4, 2013.
[16] W. Kleiminger, C. Beckel, T. Staake, and S. Santini, "Occupancy Detection from Electricity Consumption Data," *BuildSys'13*, 2014.
[17] O. Parson, S. Ghosh, M. Weal, and A. Rogers, "Using Hidden Markov Models for Iterative Non-Intrusive Appliance Modeling," *Neural Information Processing Systems Workshop on Machine Learning with Sustainability*, 2011.
[18] Z. Han, R. X. Gao, and Z. Fan, "Occupancy and Indoor Environment Quality Sensing for Smart Buildings," *Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 882–887, May 2012.
[19] B. Ai, Z. Fan, and R. X. Gao, "Occupancy Estimation for Smart Buildings by an Auto-Regressive Hidden Markov Model," *American Control Conference*, 2014.
[20] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[21] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, 1989.
[22] M. I. Jordan, *An Introduction to Probabilistic Graphical Models*. In preparation, 2007.
[23] D. B. Rubin, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.
[24] W. A. Fuller, *Introduction to Statistical Time Series*. Wiley-Interscience, 1995.