Virtual Metering: An Efficient Water Disaggregation Algorithm via Nonintrusive Load Monitoring

BINGSHENG WANG, Google Inc.
ZHIQIAN CHEN, Virginia Tech
ARNOLD P. BOEDIHARDJO, U. S. Army Corps of Engineers
CHANG-TIEN LU, Virginia Tech

The scarcity of potable water is a critical challenge in many regions around the world. Previous studies have shown that knowledge of device-level water usage can lead to significant conservation. Although there is considerable interest in determining discriminative features via sparse coding for water disaggregation to separate whole-house consumption into its component appliances, existing methods lack a mechanism for fitting coefficient distributions and are thus unable to accurately discriminate parallel devices' consumption. This article proposes a Bayesian discriminative sparse coding model, referred to as Virtual Metering (VM), for this disaggregation task. Mixture-of-Gammas is employed for the prior distribution of coefficients, contributing two benefits: (i) guaranteeing the coefficients' sparseness and non-negativity, and (ii) capturing the distribution of active coefficients. The resulting method effectively adapts the bases to aggregated consumption to facilitate discriminative learning in the proposed model, and devices' shape features are formalized and incorporated into Bayesian sparse coding to direct the learning of basis functions. Compact Gibbs Sampling (CGS) is developed to accelerate the inference process by utilizing the sparse structure of coefficients. The empirical results obtained from applying the new model to large-scale real and synthetic datasets revealed that VM significantly outperformed the benchmark methods.

CCS Concepts: • **Computing methodologies → Learning latent representations**; **Regularization**; *Feature selection*;

Additional Key Words and Phrases: Computational sustainability, Bayesian discriminative learning, sparse coding, Mixture-of-Gammas, low-sampling-rate disaggregation, non-intrusive load monitoring

ACM Reference format:
Bingsheng Wang, Zhiqian Chen, Arnold P. Boedihardjo, and Chang-Tien Lu. 2018. Virtual Metering: An Efficient Water Disaggregation Algorithm via Nonintrusive Load Monitoring. *ACM Trans. Intell. Syst. Technol.* 9, 4, Article 39 (January 2018), 30 pages.
https://doi.org/10.1145/3141770

1 INTRODUCTION

The shortage of fresh water is emerging as one of the most critical resource issues facing our society. Gilbert reported that around 80% of the world's population suffers from water shortages

B. Wang and Z. Chen contributed equally to this article.

Authors' addresses: B. Wang, 1600 Amphitheatre Parkway, Mountain View, CA 94403; email: bingsheng@google.com; Z. Chen, Room 317, Northern Virginia Center, Virginia Tech, Haycock Rd, Falls Church, VA 22043; email: czq@vt.edu; A. P. Boedihardjo, 4552 Fair Valley Dr, Fairfax VA 22033; email: Arnold.p.boedihardjo@vt.edu; C.-T. Lu, Room 310, Northern Virginia Center, Virginia Tech, Haycock Rd, Falls Church, VA 22043; email: ctlu@vt.edu.

https://doi.org/10.1145/3141770

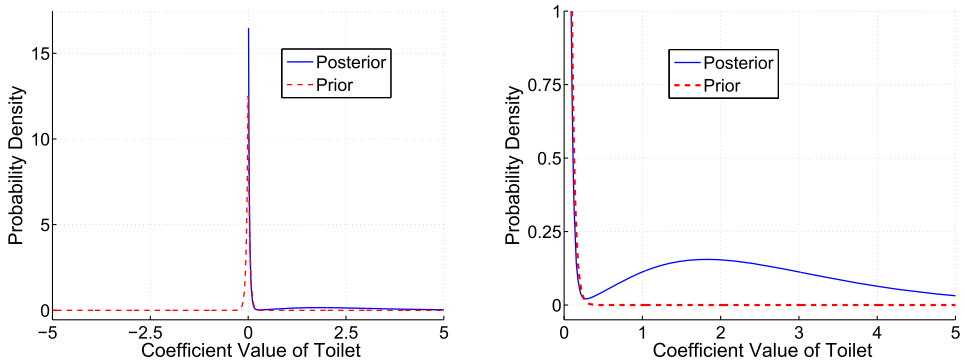

Fig. 1. The Laplace prior and smoothed posterior of Toilet's coefficients. The Laplace prior is plotted for the interval $[-5, 5]$, and the posterior is plotted for the interval $[0, 5]$ due to its non-negativeness. **Left:** Whole view of Laplace prior versus posterior. The Laplace prior can only guarantee the sparseness with the peak near zero; for the posterior, the large peak near zero is evidence that the coefficients are sparse while the small peak near 1.8 illustrates the cluster of active coefficients. **Right:** Close-up of the area near the small peak, showing the comparison of the posterior and the Laplace prior.

(Gilbert 2010), and existing resources are only sufficient to fulfill our needs for the next 60–70 years (Gleick 2000). In the United States, urban water consumption is responsible for 50–80% of the water supplied by public water supply systems and 26% of the nation's entire usage (Vickers 2001), so significant efforts are now being devoted toward residential water conservation (Chen et al. 2011; Larson et al. 2012). Previous studies have shown that detailed information on the water consumed by individual devices can help consumers reduce their consumption significantly (Froehlich et al. 2012). Water disaggregation, the task of separating total consumption into its composite components, is thus crucial for establishing a sustainable future for potable water (Gerwen et al. 2006). The research reported here therefore sought to build an effective disaggregation model capable of automatically extracting useful insights from thousands of households' smart meter readings.

Sparse coding (or sparse representation) provides excellent general models for signal classification, recognition, and source separation (Jiang et al. 2011; Roweis 2001; Virtanen 2004; Grosse et al. 2007; Mairal et al. 2009; Wright et al. 2009). Water disaggregation can conveniently be formulated as a source separation or classification problem. For the regularization term, the following issue must be considered: Does the Laplace distribution (l_1 term) overpenalize true large coefficients? For low sampling rate (1/900 Hz) sensing data, the consumption amounts for each of the devices usually combine to form a cluster which is much larger than zero. This will create at least two peaks in the distribution of coefficients: One large peak near zero represents the sparse structure, and one or more small peaks located far from zero represent the active coefficients. For example, the consumption for the device Toilet will typically be about 0.5–4.5 gallons, so with normalized basis functions, the active coefficients for Toilet will form one or more clusters which are far away from zero while its nonactive coefficients are near zero. Figure 1 showed the Laplace prior and smoothed posterior of Toilet's coefficients, revealing that the Laplace prior lacks the capability to fit the coefficients because (i) the left part shows that it is unable to inherently secure the coefficients' non-negativeness[1] and (ii) the right part indicates that the prior fails to capture the small peak of active coefficients. We therefore applied the Mixture-of-Gammas prior to mitigate these two problems.

[1]In the context of this disaggregation problem, the coefficients are not allowed to be negative since negative coefficients are meaningless for constructing device consumption components.

The key idea of Virtual Metering (VM) is to deploy and measure the device-level consumption in some homes and then use the data gathered to estimate the device-level consumption for other homes based on their aggregated smart meter readings, rendering it unnecessary to install submeters on individual devices by estimating the consumption using statistical algorithms. **Mixture-of-Gammas** can then be applied as the prior distribution for the coefficients to alleviate the overpenalization of the l_1 term due to the true large active coefficients. Compared with Laplace, Student-t or mixture-of-Gaussians (Attias 1999; Olshausen and Millman 2000), Mixture-of-Gammas will not lead to negative activations, but instead will guarantee the sparseness and capture realistic coefficient distributions: One Gamma captures those nonactive coefficients with a small shape and large rate distribution, while one or more other Gammas capture active coefficients with a small rate distribution. Next, a **sparse coding-based statistical framework** is proposed for discriminatively training the basis functions, targeted specially at accurately separating the aggregated data. After implementing a formal process to abstract the **shape features** for each device, the basis functions are initialized using the invariant features and smoothed to adapt to the variances of label training data. The parameters for each device are estimated under the Bayesian framework using Expectation Maximization (EM) and Gibbs sampling. The learned parameters for each device can then be combined for disaggregation. To enhance the new model's power of separation, the compound basis functions are further learned to adapt the bases to the aggregated consumption. By utilizing activations' sparse structure, Compact Gibbs Sampling (CGS) is designed to expedite the model's learning. In summary, the contributions of this article are as follows:

- **Design of a Bayesian discriminative disaggregation model:** Bayesian sparse coding with Mixture-of-Gammas prior is utilized as the generative model for each device. A Bayesian discriminative model is presented for disaggregation by first combining the learned models of each device, after which the discriminative capability of the combined model is enhanced through adapting bases to the aggregated data.
- **Analyses and formalizations of shape features for smart meter readings:** Rigorous analyses and definitions of shape features are presented by exploring the prior knowledge with respect to individual device consumption patterns. We use the results of the analyses to direct the learning of basis functions, and we show how it can help improve the disaggregation performance.
- **Development of efficient and effective learning algorithms:** CGS is presented for efficient inference and learning by utilizing the sparse structure of activations. CGS is capable of drawing equivalent samples to those utilized in conventional Gibbs sampling but requires on average substantially fewer operations per sample.
- **Comprehensive experiments to validate the effectiveness and efficiency:** We demonstrated the effectiveness and efficiency of the proposed model with extensive experiments based on both real and synthetic datasets. The evaluation results show that our model significantly outperformed the baselines at both the whole-home and device levels.

This article is organized as follows: Section 2 surveys related work on water disaggregation and sparse learning. Section 3 introduces the terminologies used and necessary background knowledge. Section 4 describes the sparse coding-based statistical formulation for VM. Section 5 provides the algorithms for inference and parameters estimation. The effectiveness of VM is illustrated with extensive experiments in Section 6. Section 7 presents our conclusions.

2 RELATED WORK

Energy Disaggregation and Nonintrusive Load Monitoring (NILM): With the widespread deployment of smart meters in many countries, water disaggregation is emerging as an interesting

new research direction in urban computing (Zheng et al. 2014). Pressure-based sensors have been designed for installation on water fixtures to help identify activity and estimate the corresponding attributes, such as consumption and quality, for individual household devices (Larson et al. 2012; Froehlich et al. 2009, 2011; Liu et al. 2016). By utilizing both occupancy sensors and whole-house water flow meter data, Srinivasan et al. (2011) categorized the aggregated consumption at the fine-grained device level. Although such methods are capable of achieving about 90% accuracy, they depend on high-sampling-rate sensing data (as high as 1 Khz) to capture the characteristic open–close signatures of devices. Recently, Ellert et al. (2016) modified the Viterbi algorithm to apply a supervised method to an unsupervised disaggregation problem without requiring the installation of water submeters or water sensors. A Hidden Markov Model (HMM)-based approach was developed in Chen et al. (2011) for separating low-sampling-rate (1/900 Hz) data, while (Nguyen et al. 2013a, 2013b) proposed a hybrid combination of HMM and Dynamic Time Warping (DTW) to automate the categorization of residential water end use events and estimate devices' consumption. Makonin et al. (2016) presented a new load disaggregation algorithm that ultilized a super-state hidden Markov model and a new Viterbi algorithm variant which caputures dependencies between loads. However, an HMM-based structure inherently restricts its ability to infer consumption for parallel devices. Wang et al. (2012) applied a featured discriminative dictionary to extend the sparse coding model to estimate devices' consumption from aggregated data, but this approach lacks a mechanism to capture the distribution of coefficients. A deep sparse coding-based model was presented in Dong et al. (2013) that fully utilizes the limited label data and performs disaggregation in a sequential manner; however, the model is sensitive to the disaggregation structure, and the learning process is of high computational complexity when seeking the optimal architecture for disaggregation. Recently, different deep learning models such as Recurrent Neural Network (RNN) (Kelly and Knottenbelt 2015), Convolutional Neural Network (CNN) (Kelly and Knottenbelt 2015; Zhang et al. 2016), Auto encoder (Kelly and Knottenbelt 2015), and a combination of deep learning and HMM (Huss 2012; Zhang et al. 2016; Mauch and Yang 2016) have been applied to the energy disaggregation problem.

Sparse Learning: For water disaggregation, the key is the ability to identify the discriminative features of devices; that is, to learn the discriminative basis functions for sparse coding. The convex learning form of basis functions is derived via probabilistic reasoning by maximizing the likelihood functions (Olshausen and Field 1996; Lewicki and Sejnowski 2000) and then applying an iterative refinement method for training the dictionaries (Engan et al. 1999; Kreutz-Delgado et al. 2003). Based on the given signal and dictionary, several algorithms have been proposed for the computation of the representation coefficients. Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP), and Order Recursive Matching Pursuit (ORMP) all provide solution to l_0 regularization (Pati et al. 1993; Mallat and Zhang 1993; Gharavi-Alkhansari and Huang 1998). Basis Pursuit (BP) and coordinate descent are presented for l_1 regularization (Chen et al. 1998; Wu and Lange 2008), while FOCal Underdetermined System Solver (FOCUSS) is designed for l_p $(0 < p \leq 1)$ regularization (Gorodnitsky and Rao 1997). Although an l_1, l_0, or l_p regularization term can be used to derive sparse solutions, this approach does suffer from a lack of variability for the customization of coefficients' prior distribution. The specific choice of sparse prior is critical to the objective optimization and basis function learning (Körding et al. 2003). Olshausen and Millman (2000) introduced Mixture-of-Gaussians to better match the posterior distribution by adapting the parameters of prior to the data, where Gibbs sampling is employed for sampling of intractable posterior distributions (Geman and Geman 1984). However, it lacks a mechanism to guarantee the coefficients' non-negativeness. A Laplace Scale Mixture (LSM) prior was suggested by Garrigues and Olshausen (2010) to model dependencies among coefficients, leading to multiplicative modulation and group

sparsity. Even though this method has many advantages, it requires a topographical layout for organizing the features by solving a large-scale optimization problem.

3 PRELIMINARIES

Notations and concepts: Suppose there is a total of D devices, such as Toilet and Shower. For each device $d = 1, 2, \ldots, D$, $\mathbf{Y}^{(d)} \in \mathbb{R}^{N \times P}$ is used to denote its consumption matrix, where N is the number of intervals in one day and P is the number of days. The pth day's consumption of device d is denoted as $\mathbf{y}^{(d)}_{\cdot,p}$. The water usage of device d for interval i of day p is denoted as $y^{(d)}_{i,p}$. $\bar{\mathbf{Y}}$ is used to indicate the aggregated water consumption over all devices: $\bar{\mathbf{Y}} = \sum_{d=1}^{D} \mathbf{Y}^{(d)}$. The pth column of $\bar{\mathbf{Y}}$ holds the aggregated consumption of the pth day for a given household. The ith element of $\bar{\mathbf{y}}_{\cdot,p}$, denoted as $\bar{y}_{i,p} = \sum_{d=1}^{D} y^{(d)}_{i,p}$, is the aggregated consumption at interval i in day p. During the training course, we have the individual device's consumption data, $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(D)}$, while during the testing course, only the aggregated data $\bar{\mathbf{Y}}$ are available, with the goal being to separate it into $\hat{\mathbf{Y}}^{(1)}, \hat{\mathbf{Y}}^{(2)}, \ldots, \hat{\mathbf{Y}}^{(D)}$.

Sparse coding for the disaggregation task:[2] Sparse coding techniques represent a signal $\mathbf{y} \in \mathbb{R}^{N \times 1}$ (in this case, smart meter readings) using a small number of basis functions chosen from an overcomplete dictionary $\mathbf{H} \in \mathbb{R}^{N \times M}$:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{u}, \tag{1}$$

where $\mathbf{x}$ is a sparse activation/coefficient vector belonging to $\mathbb{R}^{M \times 1}$, and $\mathbf{u} \in \mathbb{R}^{N \times 1}$ represents the noise. Intuitively, the best method to derive sparse coding coefficients $\mathbf{x}$ is to apply the l_0 constraint by optimizing the following problem:

$$\min_{\mathbf{H},\mathbf{x}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_0, \tag{2}$$

where $\|\mathbf{x}\|_0$ is the number of non-zero items in vector $\mathbf{x}$. Due to the fact that optimization with an l_0 term is NP-hard (Donoho 2006), a typical approximation is to penalize the coefficients via an l_1 regularization formulation:

$$\min_{\mathbf{H},\mathbf{x}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1. \tag{3}$$

From a Bayesian perspective, the l_1 norm is equivalent to using a Laplace prior on the coefficients $\mathbf{x}$,

$$P(\mathbf{x}) = \frac{1}{2b} \exp\left(-\frac{\|\mathbf{x}\|_1}{b}\right). \tag{4}$$

The sparse coding utilized for the disaggregation approach originates from the source separation problem (Schmidt and Olsson 2006; Schmidt et al. 2007): First train individual models for each device $\mathbf{Y}^{(d)}$ and then use the learned models to decompose an aggregated signal. Formally, the data matrix for device d is modeled as $\mathbf{Y}^{(d)} = \mathbf{H}^{(d)}\mathbf{X}^{(d)} + \mathbf{u}^{(d)}$, where $\mathbf{H}^{(d)} \in \mathbb{R}^{N \times M_d}$ is the dictionary (or basis functions) for device d, and the columns of $\mathbf{H}^{(d)}$ contain a set of M_d basis functions; $\mathbf{X}^{(d)} \in \mathbb{R}^{M_d \times P}$ represents the activations (or coefficients) of the device d's dictionary, and $\mathbf{u}$ denotes the 0-mean, τ-precision white noise (Olshausen and Field 1997). Additionally, the activations $\mathbf{X}^{(d)}$ are designated sparse (i.e., most of $\mathbf{X}^{(d)}$ are zero entries), which is designed for learning overcomplete representations of the data. As stated in the introduction, a common approach for achieving sparsity is to apply an l_1 regularization to the activations. Due to the non-negative nature of water consumption, a further constraint is adopted to ensure the non-negativeness of both activations

[2] A showcase of disaggregation is shown in A.

and bases (Hoyer 2002). Specifically, for each device d, the basis functions can be learned using a non-negative sparse coding objective:

$$\min_{\mathbf{X}^{(d)} \geq 0, \mathbf{H}^{(d)} \geq 0} \frac{1}{2} \left\| \mathbf{Y}^{(d)} - \mathbf{H}^{(d)} \mathbf{X}^{(d)} \right\|_F^2 + \lambda \sum_{q,r} (\mathbf{X}^{(d)})_{qr} \\ \text{subject to } \left\| \mathbf{H}^{(d)}_{\cdot,j} \right\|_2 \leq 1, j = 1, 2, \ldots, M_d \tag{5}$$

where $\mathbf{Y}^{(d)}$, $\mathbf{H}^{(d)}$, and $\mathbf{X}^{(d)}$ are as defined in the previous section, and $\lambda \in \mathbb{R}_+$ is a regularization parameter to balance the importance of sparseness and reconstruction error. $\|\mathbf{A}\|_F$ is the Frobenius norm, and $\|\mathbf{a}\|_2$ is the l_2 norm. This optimization problem is convex in the optimization of each variable while holding the other fixed. The objective function defined in Equartion (5) can be optimized by minimizing the objective alternatively with respect to $\mathbf{X}^{(d)}$ and $\mathbf{H}^{(d)}$.

Using the preceding procedure, the basis functions can be learned for each device $d = 1, 2, \ldots, D$ (i.e., $\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \ldots, \mathbf{H}^{(D)}$). The previously unseen aggregated signal $\bar{\mathbf{Y}}$ can be disaggregated into the D components. The individual basis functions are concatenated to form a compound basis function, and the following objective can then be used to estimate each device's activations:

$$\hat{\mathbf{X}}^{(1:D)} = \underset{\mathbf{X}^{(1:D)} \geq 0}{\operatorname{argmin}} \frac{1}{2} \left\| \bar{\mathbf{Y}} - \begin{bmatrix} \mathbf{H}^{(1)} & \cdots & \mathbf{H}^{(D)} \end{bmatrix} \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(D)} \end{bmatrix} \right\|_F^2 + \lambda \sum_{d,q,r} (\mathbf{X}^{(d)})_{qr}, \tag{6}$$

where $\mathbf{X}^{(1:D)} = [(\mathbf{X}^{(1)})^T, \ldots, (\mathbf{X}^{(D)})^T]^T$. Now we are ready to estimate device d's consumption:

$$\hat{\mathbf{Y}}^{(d)} = \mathbf{H}^{(d)} \hat{\mathbf{X}}^{(d)}. \tag{7}$$

The sparse coding model is designed to learn the reconstruction dictionaries for each device, but it lacks a mechanism to derive basis functions that can minimize the disaggregation error.

Discriminative sparse coding: The basic idea of discriminative sparse coding (Kolter et al. 2010) is to employ the regularized disaggregation error as the objective function in place of using the default non-negative sparse coding objective:

$$E_{reg} = \sum_{d=1}^{D} \frac{1}{2} \left\| \mathbf{Y}^{(d)} - \mathbf{H}^{(d)} \hat{\mathbf{A}}^{(d)} \right\|_F^2 + \lambda \sum_{d,q,r} (\hat{\mathbf{X}}^{(d)})_{qr}, \tag{8}$$

where $\hat{\mathbf{X}}^{(1:D)}$ is achieved by optimizing Equation (6). Minimizing E_{reg} is likely to achieve much better basis functions than optimizing Equation (5) for separating the aggregated signal. The best possible value of $\hat{\mathbf{X}}^{(d)}$ can be achieved by

$$\tilde{\mathbf{X}}^{(d)} = \underset{\mathbf{X}^{(d)} \geq 0}{\operatorname{argmin}} \frac{1}{2} \left\| \mathbf{Y}^{(d)} - \mathbf{H}^{(d)} \mathbf{A}^{(d)} \right\|_F^2 + \lambda \sum_{q,r} (\mathbf{X}^{(d)})_{qr}. \tag{9}$$

It is obvious that the coefficients achieved by optimizing Equation (9) are the same as the activations obtained after iteratively optimizing the non-negative sparse coding objective in Equation (5). As a result, the discriminative dictionary $\tilde{\mathbf{H}}^{(1:D)}$ can be learned by minimizing Equation (8) while making the activations as close to $\tilde{\mathbf{X}}^{(1:D)}$ as possible. Since the change of bases $\mathbf{H}^{(1:D)}$ for optimizing Equation (8) would also cause the resulting optimal coefficients to be changed, the learned discriminative basis functions (i.e., $\tilde{\mathbf{H}}^{(1:D)}$) would be different from the reconstruction bases (i.e., $\mathbf{H}^{(1:D)}$). Formally, the discriminative dictionary can be learned by optimizing the augmented

regularized disaggregation error objective:

$$\tilde{E}_{reg}\left(\mathbf{Y}^{(1:D)}, \mathbf{H}^{(1:D)}, \tilde{\mathbf{H}}^{(1:D)}\right) \equiv \sum_{d=1}^{D}\left(\frac{1}{2}\left\|\mathbf{Y}^{(d)} - \mathbf{H}^{(d)}\hat{\mathbf{X}}^{(d)}\right\|_F^2 + \lambda \sum_{q,r}(\hat{\mathbf{X}}^{(d)})_{qr}\right), \quad \text{subject to } \hat{\mathbf{X}}^{(1:D)} = \underset{\mathbf{X}^{(d)}\geq 0}{\operatorname{argmin}} \frac{1}{2}\left\|\mathbf{Y}^{(d)} - \tilde{\mathbf{H}}^{(d)}\mathbf{X}^{(d)}\right\|_F^2 + \lambda \sum_{q,r}(\mathbf{X}^{(d)})_{qr} \tag{10}$$

where $\tilde{\mathbf{H}}^{(1:D)} = [\tilde{\mathbf{H}}^{(1)}, \ldots, \tilde{\mathbf{H}}^{(D)}]$. $\mathbf{H}^{(1:D)}$ are the reconstruction bases learned from sparse coding in Equation (5), while $\tilde{\mathbf{H}}^{(1:D)}$ are the discriminative bases optimized by moving $\hat{\mathbf{X}}^{(1:D)}$ as close to $\tilde{\mathbf{X}}^{(1:D)}$ as possible. It is important to note, however, that although the discriminative sparse coding model has been developed to optimize the bases and thus decrease the overall disaggregation error, it will be unable to control the behaviors of coefficients to derive accurate disaggregation results.

4 VIRTUAL METERING: BAYESIAN DISCRIMINATIVE DISAGGREGATION MODEL

Section 4.1 describes the sparse coding with Mixture-of-Gammas prior to model the generative process, and the Bayesian discriminative model is developed in Section 4.2.

4.1 Generative Model: Bayesian Sparse Coding with Mixture-of-Gammas Prior

We apply Mixture-of-Gammas as prior of coefficients to capture the consumption amount information, and a sparse coding-based generative model is presented for each device. Without loss of generality, the labels for device and day can be removed, so let $\mathbf{y} \in \mathbb{R}^{N\times 1}$ denote one day's water consumption of a particular device, $\mathbf{H} \in \mathbb{R}^{N\times M}$ denote the basis functions, $\mathbf{u}$ denote 0-mean, and τ-precision white noise. The generative model is:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{u}. \tag{11}$$

The conditional probability of one interval's consumption is given by

$$P(y_i \mid \mathbf{x}, \tau, \mathbf{H}) = \mathcal{N}(y_i \mid \mathbf{H}_i\mathbf{x}, \tau^{-1}). \tag{12}$$

Since $y_1, y_2, \ldots, y_N$ are independent and identically distributed (i.i.d.) variables, we have

$$P(\mathbf{y} \mid \mathbf{x}, \tau, \mathbf{H}) = \prod_{i=1}^{N} P(y_i \mid \mathbf{x}, \tau). \tag{13}$$

Mixture-of-Gammas are employed to model a device's coefficients: one Gamma is used to guarantee the sparseness with a small shape and large rate distribution, while the other Gammas model the active coefficients by learning certain shape and rate values. The prior distribution over x_j is given by

$$\begin{aligned} P(x_j \mid \boldsymbol{\xi}) &= \sum_{k=1}^{K} \omega_k \text{Gamma}(x_j \mid \alpha_k, \beta_k) \\ &= \sum_{k=1}^{K} \omega_k \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} x_j^{\alpha_k - 1} e^{-\beta_k x_j} \end{aligned}. \tag{14}$$

Since $x_1, x_2, \ldots, x_M$ are i.i.d. variables, we have

$$P(\mathbf{x} \mid \boldsymbol{\xi}) = \prod_{j=1}^{M} P(x_j \mid \boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{15}$$

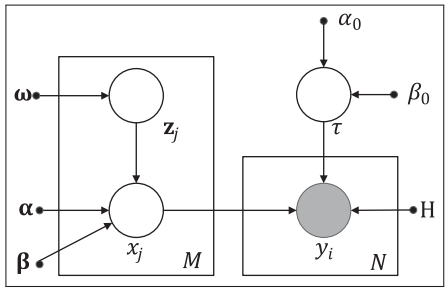

Fig. 2. Representation of the generative model as a directed acyclic graph. The observed variable y_i is shown by the shaded node, while the latent variables x_j, $\mathbf{z}_j$, and τ are represented by circles. The right box represents the N independent consumption intervals, while the left box represents the M independent coefficients. $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\mathbf{H}$ are model parameters. α_0 and β_0 are hyperparameters.

where $\boldsymbol{\xi} = \{\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ denote the parameters of coefficients, $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_K\}$, $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_K\}$, $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_K\}$, K is the number of mixtures, and ω_k denotes the mixing proportions, satisfying the condition that $\sum_{k=1}^{K} \omega_k = 1$.

The parametric form in Equation (14) provides a probabilistic generative description of the coefficient in which different Gammas play the role of hidden states. To generate x_j, a state $\mathbf{z}_j$ is first picked with probability

$$P(\mathbf{z}_j) = \prod_{k=1}^{K} (\omega_k)^{z_{jk}}. \tag{16}$$

Then x_j can be drawn from the corresponding Gamma,

$$P(x_j \mid \mathbf{z}_j) = \prod_{k=1}^{K} \left[\text{Gamma}(x_j \mid \alpha_k, \beta_k)\right]^{z_{jk}}. \tag{17}$$

Generally, given a certain state

$$P(z_{jk} = 1) = \omega_k, \tag{18}$$

it is possible to ensure the distribution to draw the coefficient:

$$P(x_j \mid z_{jk}) = \text{Gamma}(x_j \mid \alpha_k, \beta_k). \tag{19}$$

The prior distribution on τ follows a Gamma distribution with hyper-parameters α_0, β_0:

$$P(\tau \mid \alpha_0, \beta_0) = \text{Gamma}(\tau \mid \alpha_0, \beta_0). \tag{20}$$

This model can be expressed as a directed graph, illustrated in Figure 2.

4.2 Bayesian Discriminative Sparse Coding Model for Disaggregation

The generative models provided in Section 4.1 are not designed to separate the aggregated data as they lack the capability to decompose an aggregated consumption into its composite components. The generative models must be improved to adapt to the aggregated data if we are to improve disaggregation performance. Distributions of coefficients are usually invariant over time or homes since devices' consumption amplitudes will remain the same or change only slightly. A Bayesian discriminative disaggregation model is designed to promote the disaggregation performance by training the model using aggregated data while holding the parameters of coefficients' distributions fixed.

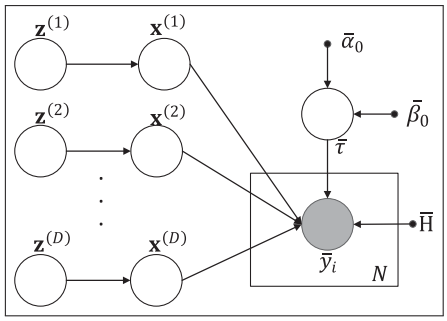

Fig. 3. Representation of the generative model for the discriminative model as a directed acyclic graph. The observed variable $\bar{y}_i$ is shown by the shaded node, while the latent variables $\mathbf{x}^{(d)}$, $\mathbf{z}^{(d)}$, and $\bar{\tau}$ are represented by circles. The box represents the N independent consumption intervals from the dataset. $\bar{\mathbf{H}}$ is the model parameter, and $\bar{\alpha}_0$ and $\bar{\beta}_0$ are hyper-parameters.

Let $\bar{\mathbf{H}} = [\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \ldots, \mathbf{H}^{(D)}]$ denote the compound basis functions. Let $\bar{\mathbf{y}} = \sum_{d=1}^{D} \mathbf{y}^{(d)}$ denote the aggregated consumption, and let

$$\bar{\mathbf{x}} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(D)} \end{bmatrix}$$

denote the compound coefficients. Our purpose is to disaggregate the total consumption $\bar{\mathbf{y}}$ into the usages of individual devices: $\hat{\mathbf{y}}^{(1)}, \hat{\mathbf{y}}^{(2)}, \ldots, \hat{\mathbf{y}}^{(D)}$. The aggregated consumption $\bar{\mathbf{y}}$ is generated by

$$\bar{\mathbf{y}} = \bar{\mathbf{H}}\bar{\mathbf{x}} + \bar{\mathbf{u}}, \tag{21}$$

where $\bar{\mathbf{u}}$ is the 0-mean, $\bar{\tau}$-precision white noise. From the generative model for the aggregated consumption, we observe that $\bar{\mathbf{x}}$ denotes the overall coefficients and $\bar{\mathbf{H}}$ denotes the basis functions for constructing the aggregated consumption. The key here is to learn the discriminative $\bar{\mathbf{H}}$ for estimating individual devices' consumption. Given the normalized basis functions, the active coefficients mainly depend on the consumption amplitude of individual devices, while the nonactive coefficients are near-zero values. The parameters of coefficients learned with individual devices' data should be optimal if they are to adequately represent the invariant patterns captured by distributions of coefficients since they are learned with device-level data. The discriminative capability of $\bar{\mathbf{H}}$ can thus be extended through training with the aggregated data while keeping parameters of coefficients unchanged (i.e., $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(D)}$).

The conditional probability of one interval's aggregated consumption is given by $\bar{y}_i$

$$P(\bar{y}_i \mid \bar{\mathbf{x}}, \bar{\mathbf{H}}, \bar{\tau}) = \mathcal{N}(\bar{y}_i \mid \bar{\mathbf{H}}_i\bar{\mathbf{x}}, \bar{\tau}^{-1}). \tag{22}$$

The aggregated interval based consumption $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_N$ are i.i.d. variables:

$$P(\bar{\mathbf{y}} \mid \bar{\mathbf{x}}, \bar{\mathbf{H}}, \bar{\tau}) = \prod_{i=1}^{N} P(\bar{y}_i \mid \bar{\mathbf{x}}, \bar{\mathbf{H}}, \bar{\tau}). \tag{23}$$

The prior distribution on $\bar{\tau}$ is given by a Gamma distribution:

$$P(\bar{\tau} \mid \bar{\alpha}_0, \bar{\beta}_0) = \text{Gamma}(\bar{\tau} \mid \bar{\alpha}_0, \bar{\beta}_0). \tag{24}$$

A graphical representation of the discriminative model is shown in Figure 3.

5 INFERENCE AND LEARNING

In this section, the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) is applied for model learning and parameters estimation. Two sampling-based inference algorithms are presented to approximate the intractable posterior. Section 5.1 introduces the inference method based on traditional Gibbs sampling (Geman and Geman 1984). To achieve more efficient and effective learning, CGS is developed in Section 5.2. The predictive density is evaluated in Section 5.3.

5.1 A Gibbs Sampler for Models' Inference

Expectation Estimation for Bayesian Sparse Coding with Mixture-of-Gammas Prior: To perform inference in the Bayesian sparse coding model, we must first construct a Gibbs sampler for generating samples from the posterior

$$P(\mathbf{x}, \mathbf{z}, \tau \mid \mathbf{y}, \theta_0, \boldsymbol{\theta}), \tag{25}$$

where $\theta_0 = \{\alpha_0, \beta_0\}$ are hyperparameters and $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{H}\}$ are model parameters. The samples of the latent variables $\mathbf{W} = \{\mathbf{x}, \mathbf{z}, \tau\}$ can be constructed by following a Gibbs sampling algorithm where we iteratively sample from the appropriate conditional distributions of $\mathbf{x}$, $\mathbf{z}$, and τ.

Given particular chosen priors, sampling τ from its conditional distribution can be reduced to a standard problem:

$$\begin{aligned} P(\tau \mid \mathbf{y}, \mathbf{z}, \mathbf{x}, \boldsymbol{\theta}) &= \frac{P(\mathbf{y}, \mathbf{z}, \mathbf{x}, \tau, \boldsymbol{\theta})}{P(\mathbf{y}, \mathbf{z}, \mathbf{x} \mid \boldsymbol{\theta})} \\ &\propto \text{Gamma}(\tau \mid \alpha_N, \beta_N) \end{aligned}, \tag{26}$$

where $\alpha_N = \alpha_0 + \frac{N}{2}$, $\beta_N = \beta_0 + \frac{1}{2}\sum_{i=1}^{N}(y_i - \mathbf{H}_i\mathbf{x})^2$. The indicator variable $\mathbf{z}$ can be sampled by sampling over individual indicators z_{jk} where $k = 1, \ldots, K, j = 1, \ldots, M$:

$$\begin{aligned} P(z_{jk} = 1 \mid \mathbf{z} \setminus z_{jk}, \mathbf{y}, \mathbf{x}, \tau, \boldsymbol{\theta}) &\propto P(z_{jk} = 1 \mid \mathbf{z} \setminus z_{jk}, x_j, \alpha_k, \beta_k, \boldsymbol{\omega}) \\ &\propto \omega_k \text{Gamma}(x_j \mid \alpha_k, \beta_k) \end{aligned}. \tag{27}$$

As expected, the value of z_{jk} mainly depends on the distribution of x_j and the value of the mixing coefficient. If ω_k is large while x_j follows a Gamma distribution with parameters α_k, β_k, then most samples of z_{jk} will be 1. The latent coefficient $\mathbf{x}$ is difficult to sample since its prior distribution is a mixture of Gammas. We can sample over the individual coefficients x_j, where $j = 1 \ldots, M$:

$$\begin{aligned} &\ln P(x_j \mid \mathbf{y}, \mathbf{x} \setminus x_j, \tau, \mathbf{z}, \boldsymbol{\theta}) \\ &\propto \sum_{i=1}^{N} -\frac{\tau}{2}\left[(H_{ij}x_j)^2 - 2H_{ij}x_j\left(y_i - \sum_{j' \neq j} H_{ij'}x_{j'}\right)\right] + \sum_{k=1}^{K} z_{jk}[(\alpha_k - 1)\ln x_j - \beta_k x_j]. \end{aligned} \tag{28}$$

The log-form of the posterior of x_j is derived given other hidden variables and observations. It is clear to see how x_j is determined. Since Equation (28) might be non-log-concave, Adaptive Rejection Metropolis Sampling (ARMS) is applied for the generation of samples (Gilks et al. 1995). Two factors in Equation (28) decide the probability value: The first factor drives the value of x_j to minimize the error $\sum_{i=1}^{N}(y_i - \mathbf{H}_i\mathbf{x})^2$, while the second factor drives x_j to follow a mixture of $K-$Gammas.

Expectation Estimation for Bayesian Discriminative Disaggregation Model: The posterior distribution of Bayesian discriminative disaggregation model is given by

$$P(\bar{\mathbf{x}}, \bar{\mathbf{z}}, \bar{\tau} \mid \bar{\mathbf{y}}, \bar{\theta}_0, \bar{\boldsymbol{\theta}}), \tag{29}$$

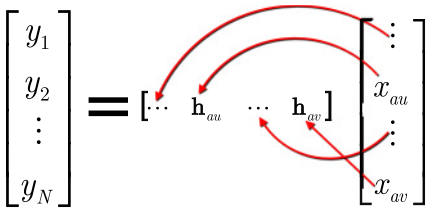

Fig. 4. Illustration of the use of sparse active coefficients. One day's consumption can be constructed with several activations of the basis functions. The lines with arrows pointing to bases denote the coefficients activating corresponding basis functions: $\ldots, x_{au}, \ldots, x_{av}$ respectively activate basis functions $\ldots, \mathbf{h}_{au}, \ldots, \mathbf{h}_{av}$, where $v \ll M$.

where $\bar{\theta}_0 = \{\bar{\alpha}_0, \bar{\beta}_0\}$ are the hyperparameters and $\bar{\theta} = \{\bar{\mathbf{H}}\}$ is the model parameter. Samples of $\bar{\tau}$ could be generated with,

$$P(\bar{\tau} \mid \bar{\mathbf{y}}, \bar{\mathbf{z}}, \bar{\mathbf{x}}, \bar{\mathbf{H}}) = \frac{P(\bar{\mathbf{y}}, \bar{\mathbf{z}}, \bar{\mathbf{x}}, \bar{\tau} \mid \bar{\mathbf{H}})}{P(\bar{\mathbf{y}}, \bar{\mathbf{z}}, \bar{\mathbf{x}} \mid \bar{\mathbf{H}})} \propto \text{Gamma}(\bar{\tau} \mid \bar{\alpha}_N, \bar{\beta}_N), \tag{30}$$

where $\bar{\alpha}_N = \bar{\alpha}_0 + \frac{N}{2}, \bar{\beta}_N = \bar{\beta}_0 + \frac{1}{2}\sum_{i=1}^{N}(\bar{y}_i - \bar{\mathbf{H}}_i\bar{\mathbf{x}})^2$. For each device $d = 1, 2, \ldots, D$, Equation (27) can be used to generate samples of $\mathbf{z}^{(d)}$, and Equation (28) can be used to generate samples of $\mathbf{x}^{(d)}$. Then samples of $\bar{\mathbf{x}}$ and $\bar{\mathbf{z}}$ are given by:

$$\begin{aligned} \bar{\mathbf{z}} &= \left[(\mathbf{z}^{(1)})^T, \ldots, (\mathbf{z}^{(M)})^T\right]^T \\ \bar{\mathbf{x}} &= \left[(\mathbf{x}^{(1)})^T, \ldots, (\mathbf{x}^{(M)})^T\right]^T. \end{aligned} \tag{31}$$

5.2 Compact Gibbs Sampling

Speeding up the estimation and model learning process is a priority. Suppose there are 10^3 days of training data, 200 basis functions for each device, and a total of 5 devices (i.e., $D = 5$). Performing 500 Gibbs iterations (i.e., $T = 500$, which is a typical number in practice) and EM can lead to convergence within 100 iterations, but, using a traditional Gibbs sampling algorithm, a total of 5×10^{10} samples of coefficients must be generated.

The key idea of CGS is to utilize the sparse structure of coefficients $\mathbf{x}$ or $\bar{\mathbf{x}}$. Taking one day's smart meter readings as an example, as shown in Figure 4, only a small number of coefficients are active while others are nonactive (i.e., near-zero values). Therefore, we can concentrate on sampling over the active coefficients while discarding the sampling process for nonactive coefficients, manually setting their values to be near zero.

Now, the problem is to predict the "identity" of coefficients for sampling; that is, to determine the candidates for active coefficients in advance of sampling. Without loss of generality, we can consider the sampling over $\mathbf{x}$ given the observation $\mathbf{y}$, model parameters $\boldsymbol{\theta}$, and other latent variables $\mathbf{z}, \tau$. The basis functions $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M]$ are overcomplete, and few bases are necessary for reconstructing the consumption data $\mathbf{y} \in \mathbb{R}^{N\times 1}$. Let $\mathbf{H}_a = [\mathbf{h}_{a1}, \ldots, \mathbf{h}_{av}]$ denote the set of basis functions which can achieve the most accurate reconstruction of $\mathbf{y}$. Note that once $\mathbf{H}_a$ is determined, the active coefficients can be identified at corresponding locations. Since it is difficult or even impossible to determine the exact $\mathbf{H}_a$, a superset of $\mathbf{H}_a$ should be identified as an alternative. Given the consumption vector $\mathbf{y}$, the qualified bases should cover at least one non-zero consumption value in $\mathbf{y}$, otherwise activating such a basis might increase the error: $E = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2$. Formally, we refer to the qualified basis for $\mathbf{y}$ as a **Candidate.**

Definition 5.1 (Candidate). Given $\mathbf{y} \in \mathbb{R}^{N\times 1}$, $\mathbf{h}_c \in \mathbf{H}$ is considered to be a **Candidate** basis function for the reconstruction of $\mathbf{y}$ if and only if the intersection between the non-zero index of elements in $\mathbf{h}_c$ and the non-zero index of elements in $\mathbf{y}$ is not empty.

Compared with the process of finding $\mathbf{H}_a$, it is relatively efficient to find the candidate bases for $\mathbf{y}$. Formally, let $\mathbf{H}_c$ denote the set of all candidate bases for $\mathbf{y}$. The relationship between $\mathbf{H}_c$ and $\mathbf{H}_a$ is given by:

Proposition 5.2. *For $\forall\ \mathbf{y} \in \mathbb{R}^{N\times 1}$, there is $\mathbf{H}_a \subseteq \mathbf{H}_c$, where $\mathbf{H}_a$ is the set of basis functions used for the most accurate reconstruction, while $\mathbf{H}_c$ contains all the candidate basis functions for the reconstruction of $\mathbf{y}$.*

Proof. Let $\mathbf{x}_a = [x_{a1}, x_{a2}, \ldots, x_{av}]^T$ denote the best coefficients used for activating the basis functions $\mathbf{H}_a$, where $x_{aq} > 0, q = 1, 2, \ldots, v$. Then the multiplication of $\mathbf{H}_a$ and $\mathbf{x}_a$ achieves the best reconstruction of $\mathbf{y}$. Thus, it achieves the minimal error:

$$E^\star = \|\mathbf{y} - \mathbf{H}_a\mathbf{x}_a\|_2^2. \tag{32}$$

Assume that $\mathbf{H}_a \nsubseteq \mathbf{H}_c$, and there is at least one basis in $\mathbf{H}_a$ but not in $\mathbf{H}_c$. Without loss of generality, let $\mathbf{h}' \in \mathbf{H}_a$ and $\mathbf{h}' \notin \mathbf{H}_c$, and $x' \in \mathbf{x}_a$ denotes the activation of $\mathbf{h}'$. Let $\mathbf{H}'_a = \mathbf{H}_a \setminus \mathbf{h}'$ and $\mathbf{x}'_a = \mathbf{x}_a \setminus x'$.

Since $x' \in \mathbf{x}_a$, then $x' > 0$. Since $\mathbf{h}' \notin \mathbf{H}_c$, no non-zero elements in $\mathbf{h}'$ overlap with the non-zero elements in $\mathbf{y}$. Thereby, $\mathbf{h}' \cdot x'$ will not contribute to the construction of $\mathbf{y}$, and

$$\begin{aligned} E &= \|\mathbf{y} - \mathbf{H}_a\mathbf{x}_a\|_2^2 \\ &= \|\mathbf{y} - \mathbf{H}'_a\mathbf{x}'_a - \mathbf{h}'x'\|_2^2 \\ &\geq \|\mathbf{y} - \mathbf{H}'_a\mathbf{x}'_a\|_2^2 + \|\mathbf{h}'x'\|_2^2 \\ &= \|\mathbf{y} - \mathbf{H}'_a\mathbf{x}'_a\|_2^2 + \sum_{i=1}^{N} (\mathbf{h}'_i x')^2 \\ &> \|\mathbf{y} - \mathbf{H}'_a\mathbf{x}'_a\|_2^2 \end{aligned} . \tag{33}$$

In Equation (33), the inequality introduced in progressing from Line 2 to 3 is based on the property of l_2 norm. Let $r_1 = \mathbf{y} - \mathbf{H}'_a\mathbf{x}'_a$ and $r_2 = \mathbf{h}'x'$. Two scenarios need to be considered:

- If there is no overlap between non-zero elements in r_1 and r_2, then the l_2 norm in Line 2 can be naturally decomposed into two parts in Line 3 and the equality holds.
- If an overlap exists between non-zero elements in r_1 and r_2, let $L_{overlap}$ denote the overlap locations in these two vectors. Then the elements in $\mathbf{y}$ at locations $L_{overlap}$ are zero since all elements in $\mathbf{h}'$ at locations $L_{overlap}$ are bigger than zero. At the overlap locations, the values of r_1 and r_2 are all negative and their l_2 norm is larger than the sum of individual l_2 norms, and the inequality holds.

The inequality in Equation (33) in Line 5 is due to $\sum_{i=1}^{N}(\mathbf{h}'_i x')^2 > 0$.

The conclusion in Equation (33) conflicts with the fact that $\mathbf{H}_a\mathbf{x}_a$ achieves the minimal error $E^\star$ in Equation (32). Therefore, the assumption is not true and $\mathbf{H}_a \subseteq \mathbf{H}_c$. □

Since $\mathbf{H}_a \subseteq \mathbf{H}_c$, then $\mathbf{H}_c$ can be considered as an alternative to $\mathbf{H}_a$. To further reduce search times for $\mathbf{H}_c$, any changes in $\mathbf{H}$ could be tracked in the maximization step: Only the basis functions whose elements change from 0 to a positive value or from a positive value to zero need to be considered for the updating of $\mathbf{H}_c$. Given $\mathbf{H}_c$, in each iteration only the corresponding coefficients $\mathbf{x}_c$ need to be sampled, and other nonactive coefficients are manually set to be near-zero values.

5.3 The Evaluation of Predictive Density

With the learned models, it is now possible to separate each device's consumption from the aggregated value. The predictive density is:

$$P\left(\hat{\mathbf{y}}^{(1)}, \hat{\mathbf{y}}^{(2)}, \ldots, \hat{\mathbf{y}}^{(D)} \mid \bar{\mathbf{y}}, \bar{\boldsymbol{\theta}}\right) = \int P\left(\hat{\mathbf{y}}^{(1)}, \hat{\mathbf{y}}^{(2)}, \ldots, \hat{\mathbf{y}}^{(D)} \mid \bar{\mathbf{W}}, \bar{\boldsymbol{\theta}}\right) P(\bar{\mathbf{W}} \mid \bar{\mathbf{y}}, \bar{\boldsymbol{\theta}}) d\bar{\mathbf{W}}$$
$$= \mathbb{E}_{\bar{\mathbf{W}}|\bar{\mathbf{y}}}\left[\prod_{d=1}^{D} P(\hat{\mathbf{y}}^{(d)} \mid \bar{\mathbf{W}}, \bar{\boldsymbol{\theta}})\right]. \tag{34}$$

Standard sampling methods can be used to achieve the expectation of $\prod_{d=1}^{D} P(\hat{\mathbf{y}}^{(d)} \mid \bar{\mathbf{W}}, \bar{\theta})$ over the posterior distribution. Suppose there is a series of samples $\bar{\mathbf{W}}^{(s+1)}, \bar{\mathbf{W}}^{(s+2)}, \ldots, \bar{\mathbf{W}}^{(T)}$ (the first s number of samples have been discarded to remove the effect of initialization), then the following equation can be evaluated

$$P\left(\hat{\mathbf{y}}^{(1)}, \hat{\mathbf{y}}^{(2)}, \ldots, \hat{\mathbf{y}}^{(D)} \mid \bar{\mathbf{y}}\right) \simeq \prod_{d=1}^{D}\left[\frac{1}{T-s}\sum_{t=s+1}^{T} P\left(\hat{\mathbf{y}}^{(d)} \mid \bar{\mathbf{W}}^{(t)}, \bar{\boldsymbol{\theta}}\right)\right]$$
$$= \prod_{d=1}^{D}\left[\frac{1}{T-s}\sum_{t=s+1}^{T}\prod_{i=1}^{N} \mathcal{N}\left(\hat{y}_i^{(d)} \mid \bar{\mathbf{H}}_i^{(d)}\bar{\mathbf{x}}^{(d,t)}, \bar{\tau}^{(t)}\right)\right]. \tag{35}$$

The mode of $\hat{y}_i^{(d)}$ is given by

$$\hat{y}_i^{(d)} = \frac{1}{T-s}\sum_{t=s+1}^{T} \bar{\mathbf{H}}_i^{(d)}\bar{\mathbf{x}}^{(d,t)}. \tag{36}$$

6 EMPIRICAL RESULTS

Comprehensive experiments on VM were conducted to evaluate the disaggregation performance. Section 6.1 introduces the experimental design and setup. Section 6.2 evaluates the scalability using synthetic datasets of various sizes. In Section 6.3, an in-depth evaluation of VM's effectiveness is performed at both the whole-home and device levels using a large scale real-world dataset. A discussion of how best to provide feedback to support consumers' water conservation efforts is presented in Section 6.4.

6.1 Dataset and Setup

Dataset: A real-world dataset was collected by Aquacraft (Mayer et al. 1999), consisting of 1,959,817 water use events recorded during a two-year study from 1,188 households across 12 study sites, including Boulder, Denver, Eugene, Seattle, San Diego, Tampa, Phoenix, Tempe and Scottsdale, Waterloo and Cambridge, Walnut Valley WD, Las Virgenes MWD, and Lompoc. Each device was labeled with one of 17 categories, and five main device types were considered in the experiments: Faucet, Dishwasher, Toilet, Shower, and Clothes Washer. This large scale dataset provides more than 30,000 days' label data for the evaluation task, and two strategies[3] have been utilized to assess the disaggregation performance: (i) **Homogeneous** evaluation is employed to verify the effectiveness and uses the same household data for both training and testing, and (ii) **heterogeneous** evaluation is applied to validate the extensibility, which trains models using some household data and tests on previously unseen households. Since the widely deployed smart meters report at a low sample rate (Chen et al. 2011), the event records were generalized into time series with a sample rate of 1/900 Hz.

[3]Detailed settings for **Homogeneous** and **Heterogeneous** evaluations are provided in Section 6.3.

Baselines: The two proposed models, Virtual Metering with CGS (VM-CGS) and Virtual Metering with Gibbs Sampling (VM-GS),[4] were compared with the following set of baselines. The first is Bayesian Sparse Coding with Mixture-of-Gammas (BSC-MoG), where the inference is performed based on Gibbs sampling, and no shape features are customized. The second baseline is Discriminative Disaggregation Sparse Coding (DDSC) (Kolter et al. 2010) with Shape Features (SF) (Wang et al. 2012); that is, DDSC+SF, which uses shape features to help the learning of DDSC's basis functions. An approach combining DDSC with its extensions Total Consumption Priors (TCP) and Group Lasso (GL) (i.e., DDSC+TCP+GL) is the third baseline. DDSC is also considered in its own right as a standalone baseline. Another baseline used for comparison is the Factorial Hidden Markov Model (FHMM) (Kim et al. 2011). We also evaluate the performance of general deep learning models by creating fully connected (FC) neural networks with typical norm constraints (l1, l2). FC-dense is referred to as fully connected neural networks without any constraint. FC-l1 is actually sparse coding using deep learning (Singh et al. 2016). All tests ran on an Intel i7-2760QM and Nvidia GTX 1070.

Evaluation Metrics: Both whole-home and device-level evaluation metrics are inspected, and the whole-home level disaggregation capability is measured utilizing Accuracy (Kolter et al. 2010) and Normalized Disaggregation Error (NDE) (Kolter and Jaakkola 2012): Accuracy evaluates the total-day accuracy of the estimation methods, while NDE measures how well the models separate individual devices' consumption from the aggregated consumption

$$\text{Accuracy} = \frac{\sum_{d,p} \min\left(\left\|\mathbf{y}_{\cdot,p}^{(d)}\right\|_1, \left\|\hat{\mathbf{y}}_{\cdot,p}^{(d)}\right\|_1\right)}{\sum_{i,p}(\bar{y})_{i,p}} \tag{37}$$

$$\text{NDE} = \sqrt{\sum_{d,p}\left(\frac{\left\|\mathbf{y}_{\cdot,p}^{(d)} - \hat{\mathbf{y}}_{\cdot,p}^{(d)}\right\|_2^2}{\left\|\mathbf{y}_{\cdot,p}^{(d)}\right\|_2^2}\right)}, \tag{38}$$

where $\hat{\mathbf{y}}_{\cdot,p}^{(d)}$ is the estimated consumption for device d at the pth day.

With respect to device-level evaluation, the quantitative precision, recall, and F-measure are applied: The precision is the fraction of disaggregated consumption that is correctly separated, recall is the fraction of true device level consumption that is successfully separated, and the F-measure for device d is: $F(d) = 2 \times \frac{\text{Precision}(d) \times \text{Recall}(d)}{\text{Precision}(d) + \text{Recall}(d)}$, where

$$\text{Precision}(d) = \frac{\sum_{i,p} \min\left(y_{i,p}^{(d)}, \hat{y}_{i,p}^{(d)}\right)}{\sum_{p,i} \hat{y}_{i,p}^{(d)}} \tag{39}$$

$$\text{Recall}(d) = \frac{\sum_{i,p} \min\left(y_{i,p}^{(d)}, \hat{y}_{i,p}^{(d)}\right)}{\sum_{p,i} y_{i,p}^{(d)}}, \tag{40}$$

where $\hat{y}_{i,p}^{(d)}$ is the estimated consumption for device d at the ith interval in the pth day. Additionally, the average F-measure is used to evaluate the models' overall disaggregation performance: $AF = \frac{1}{D}\sum_{d=1}^{D} F(d)$.

[4]The basis functions of VM-CGS and VM-GS are initialized with the help of Shape Features (SF), which was proposed in our preliminary work (Wang et al. 2012).

Table 1. Scalable Evaluations of Disaggregation Methods on **Synthetic** Data for Time Periods Varying from 50 to 1,000 Days

Methods \ Days	50	100	400	800	1000
VM-CGS	0.4625±0.0168	0.5060±0.0178	0.5965±0.0545	0.6295±0.0535	0.6532±0.0557
	0.7511±0.0851	0.7895±0.0354	**0.8561±0.0811**	**0.8971±0.0135**	0.9084±0.0524
	0.7610±0.0556	0.7246±0.0659	0.6491±0.0515	0.6062±0.0295	0.5925±0.0301
VM-GS	0.4902 ±0.0839	0.5013±0.0711	0.6060±0.0133	0.6465±0.0871	0.6709±0.0345
	0.7866±0.0566	0.7807±0.0218	0.8515±0.0520	0.8937±0.0283	**0.9158±0.0393**
	0.7589±0.0407	0.7167±0.0343	0.6205±0.0373	0.6015±0.0340	0.5872±0.0413
BSC-MoG	0.2790±0.0255	0.3597±0.0343	0.4893±0.0192	0.4925±0.0575	0.5621±0.0619
	0.6354±0.0317	0.7071±0.0635	0.7821±0.0648	0.7977±0.0243	0.8172±0.0357
	0.8346±0.0468	0.7864±0.0941	0.6967±0.0435	0.6861±0.0864	0.6429±0.0542
DDSC+SF	0.2747±0.0581	0.3235±0.0481	0.4389±0.0291	0.4704±0.0640	0.5167±0.0679
	0.6050±0.0512	0.6702±0.0376	0.7467±0.0171	0.7551±0.0115	0.7866±0.0115
	0.8702±0.0541	0.8209±0.0650	0.7593±0.0720	0.7289±0.0805	0.7091±0.0273
DDSC+TCP+GL	0.1515±0.0895	0.2186±0.0711	0.3421±0.0844	0.3858±0.0865	0.4187±0.0652
	0.4804±0.0472	0.5856±0.0333	0.6916±0.0278	0.7218±0.0217	0.7542±0.0481
	0.9110±0.0850	0.8674±0.0730	0.7839±0.0401	0.7579±0.0401	0.7289±0.0617
DDSC	0.1268±0.0931	0.2063±0.0264	0.2621±0.0275	0.3261±0.0753	0.3711±0.0357
	0.4408±0.0258	0.5523±0.0590	0.6544±0.0417	0.6739±0.0271	0.7154±0.0568
	0.9271±0.1080	0.8867±0.0692	0.7923±0.0579	0.7631±0.0401	0.7287±0.0097
FHMM	0.2843±0.0791	0.3331±0.0694	0.4277±0.0313	0.4641±0.0544	0.4801±0.0484
	0.5804±0.0631	0.6324±0.0632	0.7289±0.0396	0.7313±0.0235	0.7619±0.0135
	0.8729±0.0737	0.8134±0.0564	0.7338±0.0589	0.7128±0.0576	0.6784±0.0160
FC-l1	0.7167±0.0438	0.7280±0.0432	0.7627±0.0045	0.7634±0.0029	**0.7601±0.0025**
	0.8065±0.0533	0.8052±0.0530	0.8276±0.0027	0.8275±0.0018	0.8273±0.0011
	0.6350±0.0222	0.5624±0.0176	0.4804±0.0103	0.4792±0.0109	0.5032±0.0140
FC-l2	**0.7064±0.0419**	**0.7397±0.0405**	**0.7652±0.0236**	**0.7676±0.0049**	0.7569±0.0039
	0.7995±0.0626	0.8138±0.0472	0.8165±0.0293	0.8295±0.0025	0.8333±0.0061
	0.5659±0.0772	**0.4828±0.0825**	**0.4144±0.0664**	**0.4295±0.0508**	0.4848±0.0224
FC-dense	0.6748±0.0665	0.7056±0.0645	0.7582±0.0348	0.7611±0.0215	0.7569±0.0189
	0.8151±0.0613	**0.8265±0.0465**	0.8205±0.0266	0.8311±0.0063	0.8300±0.0107
	0.6049±0.0846	0.5491±0.1158	0.4357±0.0856	0.4493±0.0689	**0.4746±0.0538**

For each size of data, 10-fold cross-validation was applied and the mean±std of Avg. F-measure, Accuracy, and NDE are reported, where each metric occupies one line. Bold entries denote the best performance values.

6.2 Scalable Evaluation with Synthetic Data

Various sizes of synthetic data were generated (the detailed generation process is shown in B) and then used to evaluate the scalability of the proposed models (VM-CGS and VM-GS), reporting both the whole-home level performance and their comparisons with the baselines. All the methods were evaluated using the synthetic datasets, varying from 50 to 1,000 days. For each data size, 10-fold cross-validation is applied and the mean±std of Avg. F-measure, Accuracy, and NDE are shown in Table 1. The performance of our proposed models was better than that of the baselines for every data size. The performance of all the methods increased with the data size from 50 to 1,000 days, and there is a relatively large performance gap from 100 to 400 days. Regardless of the data size, VM-CGS proved to be capable of achieving nearly the same results as VM-GS, thus validating the

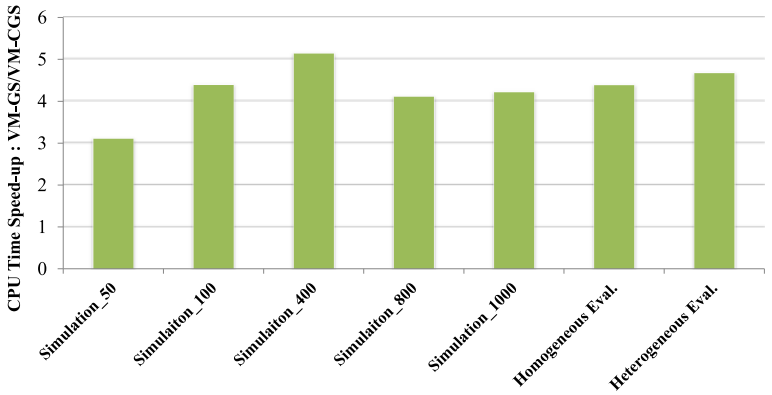

Fig. 5. Relative computation speeds achieved by VM-CGS compared to VM-GS for both simulated and real-world datasets (from left to right): Simulation data with 50 days, 100 days, 400 days, 800 days, and 1,000 days duration; homogeneous and heterogeneous evaluations with real data.

effectiveness and correctness of the proposed CGS sampling method. In comparison to BSC-MoG, VM-CGS exhibited better performances for Avg. F-measure (from 9% to 18%), Accuracy (from 7% to 12%), and NDE (from 5% to 8%), while DDSC+SF outperformed DDSC with respect to all the three metrics. These gains can be attributed to the customizations of basis functions with the help of shape features. BSC-MoG achieved higher Avg. F-measure, Accuracy, and lower NDE than DDSC, verifying that the choice of Mixture-of-Gammas was a suitable prior for the sparse coefficients. Comparing DDSC+TCP+GL with DDSC, the TCP and GL extensions played only a small part in performance enhancement. Also, FHMM achieved similar results to DDSC+SF. Deep models performed better than the other methods, especially when data are smaller. As data size increased, our proposed method significantly outperformed FC group in terms of accuracy: VM-CGS (0.8971) improved FC-l2 (0.8295) by 7% under days = 800; while VM-GS (0.9158) is 7% better than FC-dense (0.83), which is second best.

Our summary of the computation speeds achieved for the five different-sized groups of data are shown as the first five bars (from Simulation_50 to Simulation_1000) in Figure 5. Each bar shows the speedup of VM-CGS relative to VM-GS, and the consistent advantage enjoyed by VM-CGS (speedup factors from 3.1 to 5.1) clearly demonstrates that CGS considerably reduced the time required for model learning.

6.3 Water Disaggregation Using Large-scale Real World Data

To assess the disaggregation performance, the proposed models and the baselines were evaluated using two different procedures: **homogeneous** and **heterogeneous** evaluations. In the **homogeneous** evaluation, the performance was evaluated using a 10-fold cross-validation for each home: The water use events were divided for each home randomly into 10 approximately equal-sized groups; the preceding methods were trained on the combined data from nine of these groups and then tested on the group of data withheld; this was repeated, withholding each of the 10 groups in turn; and the average performance over all homes was reported. In the **heterogeneous** evaluation, to simulate a more practical and realistic situation, a 10-fold cross-validation was used across homes: All the homes were randomly divided into 10 approximately equal-sized groups and the models were trained on the combined data from nine of these groups, then the data for the

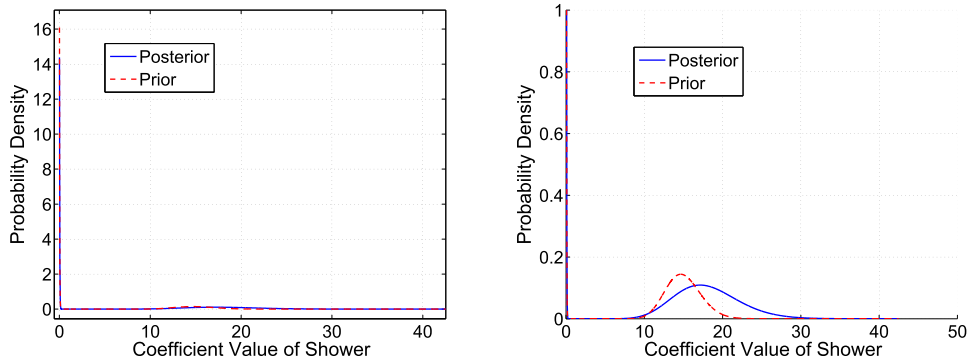

Fig. 6. A mixture of two Gammas prior and smoothed posterior of Shower's coefficients. **Left:** Whole view of prior versus posterior. Both prior and posterior have two peaks: A large peak near zero indicates the sparse structure of the activations, and the small peak near 15 indicates a cluster of active coefficients. **Right:** Close-up of the area near the small peak showing the comparison of the prior and posterior.

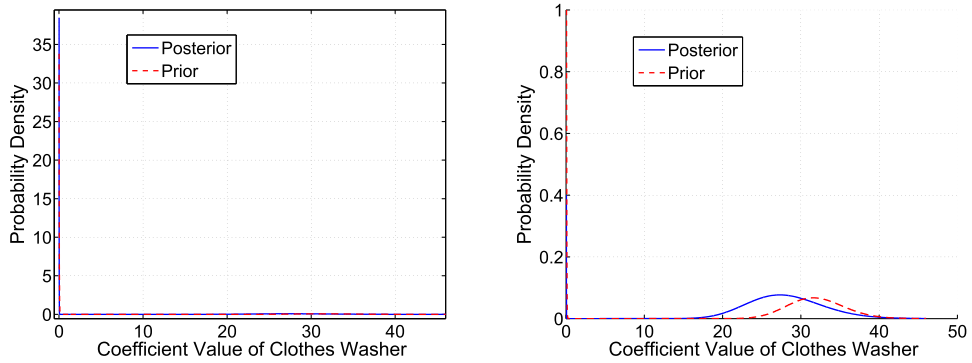

Fig. 7. A mixture of two Gammas prior and smoothed posterior of Clothes Washer's coefficients. **Left:** Whole view of prior versus posterior. Both prior and posterior have two peaks: A large peak near zero indicates the sparse structure of the activations, and the small peak near 28 indicates a cluster of active coefficients. **Right:** Close-up of the area near the small peak showing the comparison of the prior and posterior.

previously unseen group of homes were used for testing. This process was repeated, withholding the data for each of the 10 groups in turn, and the average performance was reported.

Before beginning the discussion of the disaggregation performance, let us first examine the coefficients' posterior distribution learned with **VM-CGS** under the setting of **heterogeneous** evaluations. The prior versus posterior for Shower and Washer are shown in Figure 6 and Figure 7, respectively. (The prior versus posterior for the other three devices are shown in B.2). The left part in Figures 6 and 7 shows the whole view of the prior versus posterior: In each case, the posterior had a large peak near zero, capturing the sparse nonactive coefficients' distribution. As these figures show, Mixture-of-Gammas prior was capable of capturing the large peak near zero, indicating that the customized prior was indeed effective in deriving the sparse coefficients. The small peaks in Figures 6 and 7 were used for fitting the distribution of active coefficients. The right parts of these two figures show close-ups comparing the shapes of the prior versus posterior for each, zooming in on the local region near the small peaks: The Mixture-of-Gammas prior generally captured the posterior well in spite of minor shift in the peak position and a small change in the peak scale.

Table 2. Disaggregation Results for **Homogeneous** Evaluations

Methods \ Devices	Faucet	Dishwasher	Toilet	Shower	Clothes Washer
VM-CGS	0.6159±0.1087	0.3632±0.0573	0.6802±0.0611	0.7065±0.0820	0.7352±0.1220
	0.5023±0.0311	**0.5517±0.0974**	0.5712±0.0859	**0.5925±0.0866**	0.7761±0.0813
	0.5489±0.0345	0.4335±0.0414	0.6151±0.0238	**0.6443±0.0852**	0.7525±0.0937
VM-GS	0.6187±0.0993	**0.3816±0.0266**	0.6908±0.0730	0.6946±0.1024	**0.7663±0.1035**
	0.5138±0.0332	0.5494±0.0617	0.5516±0.1145	0.5752±0.0681	**0.7856±0.0693**
	0.5574±0.0326	**0.4483±0.0171**	0.6035±0.0446	0.6292±0.0828	**0.7740±0.0782**
BSC-MoG	0.4496±0.1292	0.2121±0.0462	0.4581±0.0506	0.6493±0.0852	0.6466±0.1273
	0.4680±0.0252	0.4018±0.0526	0.4419±0.1069	0.5304±0.0556	0.7277±0.0661
	0.4522±0.0741	0.2743±0.0356	0.4468±0.0733	0.5819±0.0558	0.6823±0.0952
DDSC+SF	0.4236±0.0736	0.1555±0.0130	0.4958±0.1072	0.6851±0.0980	0.6230±0.0790
	0.3268±0.1788	0.4728±0.0914	0.6092±0.0934	0.5185±0.1345	0.4847±0.0728
	0.3347±0.1208	0.2324±0.0018	0.5363±0.0385	0.5756±0.0432	0.5451±0.0760
DDSC+TCP+GL	0.2686±0.0567	0.1217±0.0440	0.4995±0.0787	0.3753±0.0776	0.3887±0.0674
	0.5472±0.0738	0.3064±0.0335	0.2317±0.0414	0.5472±0.0738	0.3682±0.0973
	0.3570±0.0559	0.1722±0.0475	0.3142±0.0421	0.4370±0.0269	0.3750±0.0778
DDSC	0.2614±0.0582	0.1152±0.0522	0.4603±0.1303	0.3613±0.1213	0.3572±0.0940
	0.4572±0.1351	0.2764±0.0456	0.2095±0.0788	0.6313±0.0718	0.4394±0.0568
	0.3187±0.0057	0.1557±0.0451	0.2759±0.0778	0.4563±0.1153	0.3890±0.0633
FHMM	0.3626±0.1008	0.1953±0.0768	0.4256±0.0181	0.4808±0.1486	0.5619±0.1157
	0.4720±0.0543	0.4865±0.0687	0.6646±0.0711	0.2828±0.0593	0.4002±0.1122
	0.4001±0.0551	0.2720±0.0850	0.5185±0.0342	0.3547±0.0869	0.4663±0.1148
FC-l1	0.8137±0.0170	0.0683±0.1043	0.8466±0.0154	0.7282±0.0090	0.4331±0.0005
	0.7249±0.0239	0.0061±0.0093	**0.6673±0.0280**	0.3691±0.0191	0.0617±0.0083
	0.7662±0.0061	nan±nan	**0.7457±0.0120**	0.4894±0.0148	0.1078±0.0125
FC-l2	0.8396±0.0291	0.0341±0.0812	0.8642±0.0213	0.7349±0.0094	0.4332±0.0007
	0.6734±0.0568	0.0030±0.0072	0.6206±0.0530	0.3312±0.0412	0.0536±0.0104
	0.7448±0.0240	nan±nan	0.7203±0.0286	0.4548±0.0373	0.0950±0.0161
FC-dense	**0.8496±0.0375**	0.0455±0.0910	**0.8699±0.0313**	**0.7407±0.0132**	0.4337±0.0008
	0.6522±0.0763	0.0038±0.0077	0.6097±0.0733	0.3118±0.0596	0.0464±0.0154
	0.7332±0.0363	nan±nan	0.7127±0.0400	0.4348±0.0585	0.0830±0.0254

Ten-fold cross-validation was applied for each home, and the mean±std of precision, recall, and F-measure are reported, where each metric occupies one line. The bold entries denote the best F-measure.

Homogeneous Evaluation Results: The device and whole-home level evaluation results under the **homogeneous** setting are shown in Table 2 and Figure 8, respectively. At the device level, VM-CGS was capable of achieving similar results to those obtained using VM-GS regarding qualitative Precision, Recall, and F-measure, proving that CGS could indeed generate samples that were as good as Gibbs sampling (GS). VM-CGS provided better performance than BSC-MoG, with 6% to 22% improvements in Precision, 3% to 15% improvements in Recall, and 6% to 17% improvements in F-measure. Meanwhile, DDSC+SF outperformed DDSC. These findings indicate that shape features were helpful for improving performance. The fact that BSC-MoG achieved higher Precision, Recall, and F-measure than DDSC showed that the Mixture-of-Gamma prior was once again a good choice. TCP and GL were of little significance in performance enhancement since the performance of DDSC+TCP+GL was similar to that of DDSC. Although FHMM could achieve acceptable results,

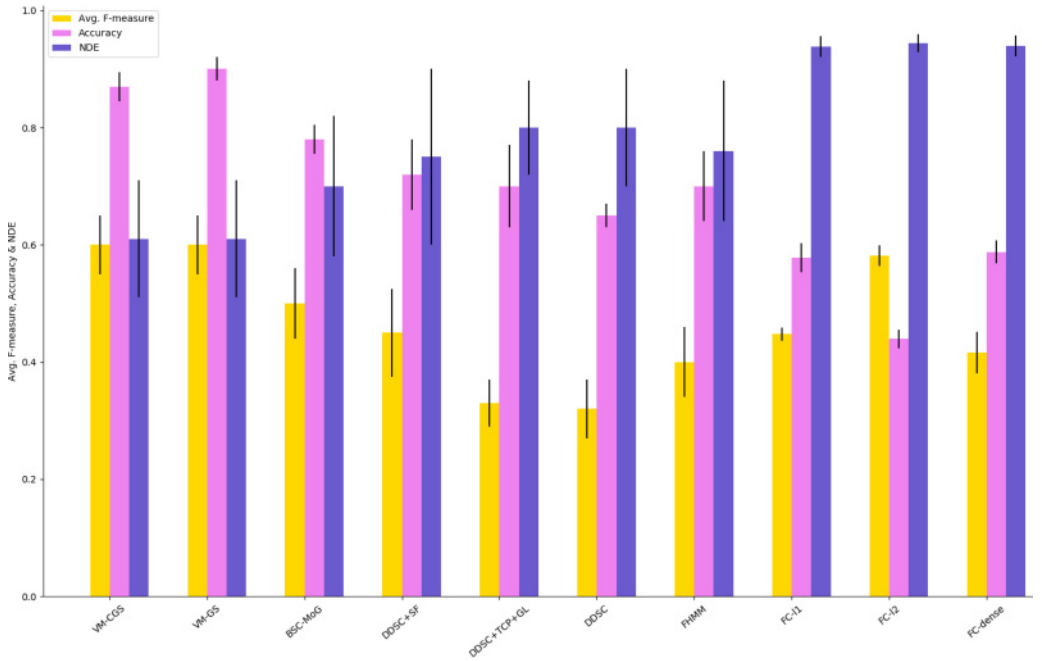

Fig. 8. Whole-home level performance report for **homogeneous** evaluations: 10-fold cross-validation was applied and the mean±std of Avg. F-measure, Accuracy, and NDE are plotted as bars.

very similar to those produced by DDSC+SF. At the whole-home level, as expected, the values of Avg. F-measure, Accuracy, and NDE achieved by VM-CGS were almost the same as those achieved by VM-GS. Employing shape features enabled VM-CGS and DDSC+SF to outperform BSC-MoG and DDSC, respectively. The function of Mixture-of-Gammas prior allowed BSC-MoG to produce higher Avg. F-measure and Accuracy, and lower NDE than DDSC. DDSC+TCP+GL exhibited a similar performance to DDSC, while FHMM had a similar performance to DDSC+SF. The FC group dominated the performance on Faucet and Toilet. However, it performed extremely badly with Dishwasher and Clothes Washer. This is because the distributions for Dishwasher and Clothes washer are different from the others, as shown in Figure 9. The distributions of Faucet and Toilet were smooth, while that of Clothes Washer and Dish Washer were not smooth. Although deep learning enjoys the advantage of neural networks that can model hidden functions, non-smooth function poses a challenge to deep modeling.

Heterogeneous Evaluation Results: The device and whole-home level evaluation results are shown in Table 3 and Figure 10, respectively. With respect to the device-level performance, there was little difference between the performance of VM-CGS and VM-GS, which confirmed that CGS was as effective as GS. Table 3 also demonstrates that VM-CGS and DDSC+SF, respectively, improved the device-level performance of BSC-MoG and DDSC with respect to precision, recall, and F-measure, illustrating the effectiveness of shape features in the learning of models. BSC-MoG outperformed DDSC, validating the decision to use Mixture-of-Gammas for the learning of coefficients. The performance of DDSC+TCP+GL was similar to that of DDSC, demonstrating that TCP and GL contributed little to performance improvement. FHMM provided barely satisfactory performance, similar to that of DDSC+SF. For the whole-home level performance evaluation results, the performance of VM-CGS was almost equivalent to that of VM-GS with respect to Avg.

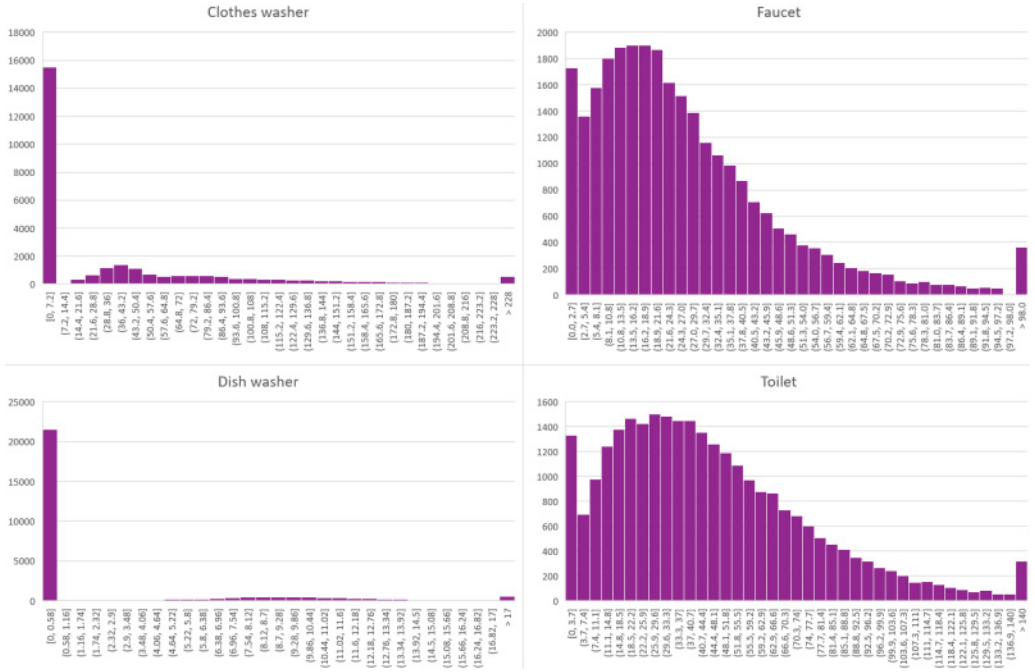

Fig. 9. Distributions difference among Clothes Washer, Dish Washer, Faucet, and Toilet.

F-measure, Accuracy, and NDE. By using shape features, VM-CGS and DDSC+SF, respectively, outperformed BSC-MoG and DDSC. Comparing the results for BSC-MoG and DDSC in terms of Avg. F-measure, Accuracy, and NDE confirmed that Mixture-of-Gammas prior was indeed a correct choice for modeling the coefficients' distribution. As the data presented in the table shows, DDSC performed as well as DDSC+TCP+GL and FHMM performed as well as DDSC+SF. Similar to its performance on homogenous data, the FC group performed best on the Faucet and Toilet evaluation, but the performance dramatically declined at Dishwasher and Clothes Washer.

Cross-Comparisons: Moving on to the results of the comparison between the **homogeneous** results in Table 2 and the **heterogeneous** results in Table 3, both VM-CGS and VM-GS significantly outperformed the baseline methods in terms of qualitative precision, recall, and F-measure. As expected, the performance of all the methods in the **homogeneous** evaluations is much better than that of their corresponding **heterogeneous** evaluations, although both VM-CGS and VM-GS still achieved acceptable disaggregation results for the **heterogeneous** case. For both the homogeneous and heterogeneous cases, the baselines, especially DDSC, and DDSC+TCP+GL, performed particularly poorly on the estimation of Dishwasher's consumption, largely due to the difficulty in distinguishing the Dishwasher's consumption from the aggregated value without considering the shape features or the customization of coefficients' prior distribution. Mixture-of-Gamma greatly improved the model performance, illustrating the importance of proper priors for the learning of coefficients' representation. As shown in both Figure 8 and Figure 10, the overall performance of VM-CGS was comparable with that of VM-GS, and both considerately outperformed the baselines with respect to Avg. F-measure, Accuracy, and NDE. For all methods, the performance decreased markedly when changing from homogeneous to heterogeneous evaluations.

The greater computational speed achieved by VM-CGS over VM-GS for both homogeneous and heterogeneous evaluations are respectively illustrated by the last two bars in Figure 5; the height

Table 3. Disaggregation Results for **Heterogeneous** Evaluations

Methods \ Devices	Faucet	Dishwasher	Toilet	Shower	Clothes Washer
VM-CGS	0.4679±0.0807	0.2372±0.0686	0.6258±0.0710	0.5408±0.1194	**0.5836±0.0573**
	0.3999±0.0882	0.3900±0.1244	0.4962±0.0222	0.4185±0.0513	0.7150±0.0587
	0.4211±0.0176	0.2837±0.0659	0.5514±0.0200	0.4667±0.0543	0.6393±0.0107
VM-GS	0.4524±0.0863	**0.2541±0.0871**	0.6469±0.0404	0.5740±0.1360	0.6184±0.0590
	0.3902±0.0965	0.4193±0.0790	0.4425±0.0759	0.4362±0.0823	**0.7566±0.1034**
	0.4090±0.0374	**0.3069±0.0687**	0.5208±0.0437	0.4853±0.0702	**0.6773±0.0566**
BSC-MoG	0.3062±0.0913	0.2063±0.0800	0.3264±0.0742	0.4779±0.0732	0.4847±0.0983
	0.2932±0.0855	0.3015±0.0828	0.3454±0.0517	0.3725±0.0565	0.4082±0.0398
	0.2822±0.0290	0.2314±0.0484	0.3353±0.0636	0.4183±0.0613	0.4414±0.0615
DDSC+SF	0.2997±0.0413	0.1411±0.0334	0.3494±0.0310	0.5185±0.1345	0.5099±0.0323
	0.2435±0.1505	0.3749±0.0714	0.4299±0.0633	0.3956±0.0916	0.2719±0.0467
	0.2462±0.0711	0.2049±0.0459	0.3821±0.0177	0.4362±0.0433	0.3536±0.0452
DDSC+TCP+GL	0.1917±0.0205	0.1125±0.0081	0.1987±0.0266	0.3024±0.0549	0.2927±0.0461
	0.2807±0.1444	0.1717±0.0269	0.1716±0.0851	0.4592±0.1321	0.2500±0.1240
	0.2192±0.0460	0.1357±0.0140	0.1802±0.0590	0.3551±0.0473	0.2583±0.0863
DDSC	0.2100±0.0315	0.1110±0.0064	0.1877±0.0327	0.3011±0.0817	0.2782±0.0521
	0.2299±0.1974	0.1280±0.0472	0.1564±0.0688	**0.5658±0.1811**	0.3300±0.1465
	0.1901±0.0839	0.1158±0.0160	0.1651±0.0407	0.3719±0.0258	0.2936±0.0865
FHMM	0.3446±0.0320	0.1291±0.0199	0.3863±0.0723	0.2344±0.0233	0.4952±0.1264
	0.4601±0.0284	**0.4623±0.0527**	0.4256±0.0181	0.1649±0.0741	0.2252±0.1540
	0.3927±0.0133	0.2003±0.0209	0.4025±0.0373	0.1887±0.0520	0.3019±0.1658
FC-l1	0.7999±0.0169	0.0000±0.0000	0.8295±0.0149	0.7120±0.0107	0.4412±0.0009
	0.7163±0.0250	0.0000±0.0000	**0.6575±0.0235**	0.3743±0.0222	0.0759±0.0096
	0.7552±0.0073	nan±nan	**0.7331±0.0092**	**0.4901±0.0172**	0.1293±0.0142
FC-l2	0.8320±0.0346	0.0000±0.0000	0.8538±0.0269	0.7256±0.0160	0.4421±0.0011
	0.6536±0.0658	0.0000±0.0000	0.6041±0.0569	0.3252±0.0523	0.0606±0.0171
	0.7285±0.0277	nan±nan	0.7049±0.0298	0.4459±0.0466	0.1057±0.0262
FC-dense	**0.8553±0.0439**	0.0000±0.0000	**0.8712±0.0333**	**0.7327±0.0166**	0.4425±0.0011
	0.5904±0.1089	0.0000±0.0000	0.5466±0.0977	0.2847±0.0727	0.0500±0.0206
	0.6891±0.0655	nan±nan	0.6642±0.0673	0.4038±0.0727	0.0885±0.0327

Ten-fold cross-validation was applied, and the mean±std of precision, recall, and F-measure are reported, where each metric occupies one line. The bold entries denote the best F-measure.

of each bar shows the factor by which the average running time of VM-CGS is faster than that of VM-GS, indicating that the speed advantage of the heterogeneous evaluations was slightly better than that of the homogeneous evaluations, while VM-CGS achieved a consistent 4.4× to 4.7× improvement in speed, which was nontrivial for large real-world computations.

6.4 Water Conservation Based on Detailed Feedback

The eventual goal of water disaggregation efforts is to provide feedback to consumers to support their water conservation. VM-CGS could therefore be applied to estimate device-level consumption and provide detailed usage information to consumers.

Figure 11 illustrates the daily actual and VM-CGS's estimated consumption for an example test home under heterogeneous evaluation conditions. With respect to the Aggregated Consumption,

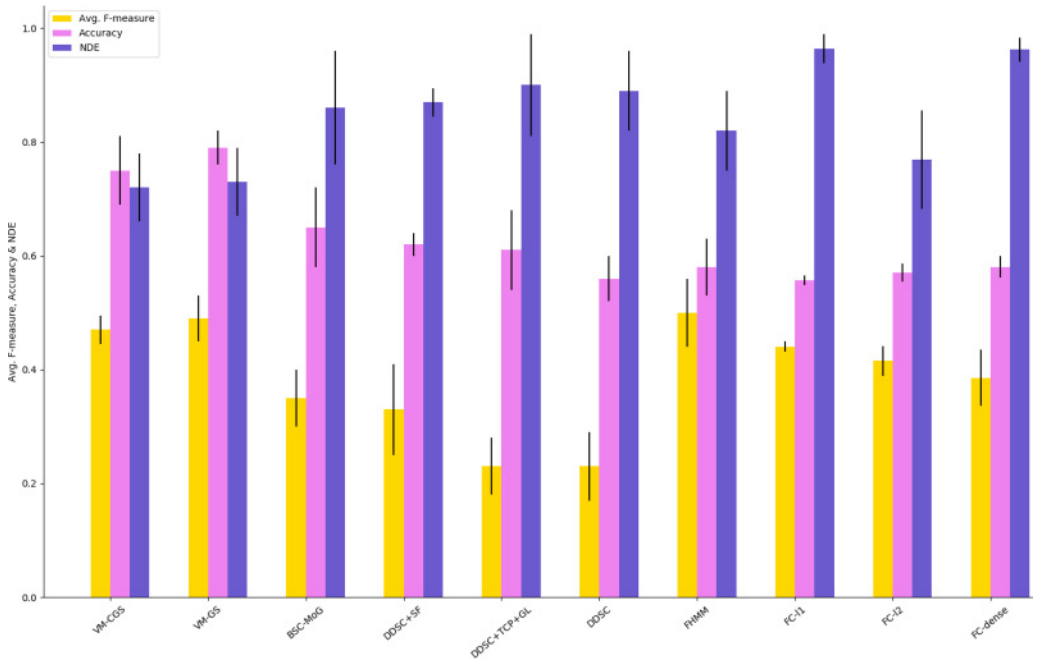

Fig. 10. Whole-home level performance report for **heterogeneous** evaluations: Ten-fold cross-validation was applied and the mean±std of Avg. F-measure, Accuracy, and NDE are plotted as bars.

VM-CGS captured most of the consumptions correctly. Despite some instances of underestimation or overestimation, the consumption of three devices, Toilet, Shower, and Clothes Washer, were all estimated fairly accurately. However, VM-CGS had problems estimating some device consumption precisely. For example, it failed to detect Faucet's usage at some points and severely overestimated Dishwasher's consumption during some intervals. Another incorrectly detected disaggregation case was that during the interval from 16:30 to 17:15, some of Shower's consumption was misdisaggregated and attributed to one of the other three devices' usage, while Faucet was overestimated at interval 16:30 and underestimated at interval 16:45. Although these instances of poor estimation cases decreased the model's performance, the estimated time series were still informative and provided useful device-level consumption patterns that could be used to inform water-saving actions.

Figure 12 shows the total actual and VM-CGS's estimated consumption percentages for all the test homes in the **heterogeneous** evaluations. VM-CGS correctly identified individual devices' percentage usage information, in spite of some overestimations for Dishwasher and Clothes Washer and slight underestimations on Faucet, Toilet, and Shower. In response to such information, effective measures can be taken for conservation, such as choosing a water-efficient shower head or buying a water-saving washing machine.

6.5 Discussion

The preceding experimental results demonstrate that VM-CGS and VM-GS significantly outperformed the baselines for the task of water disaggregation under both homogeneous and heterogeneous settings. The disaggregation performance of VM-CGS was similar to that of VM-GS, but VM-CGS achieved greater computational speed over VM-GS. The experimental results verified

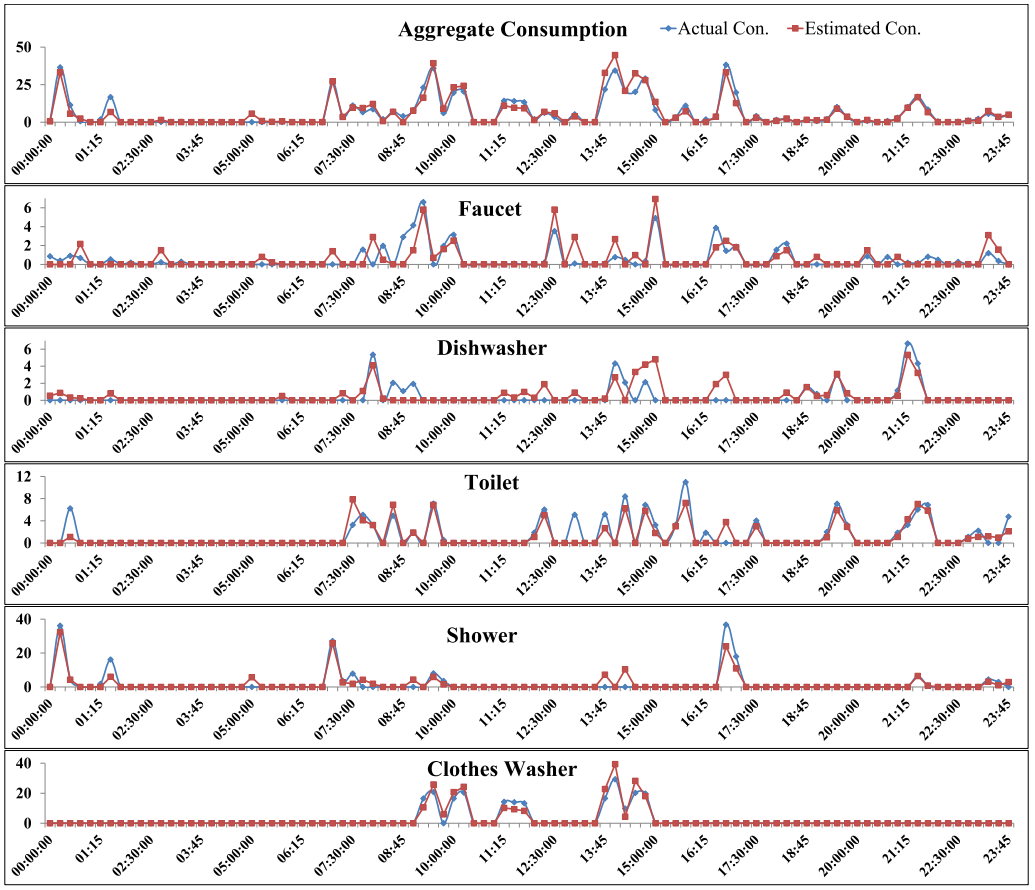

Fig. 11. Illustration of Actual Consumption (Actual Con.) and VM-CGS's Estimated Consumption (Estimated Con.) for an example home for one 24-hour period, in units of Gallons.

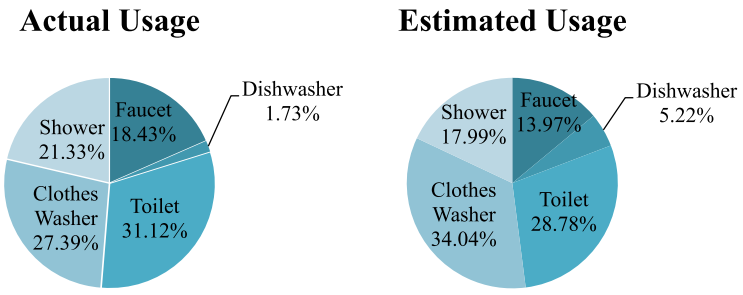

Fig. 12. Total water consumption percentages for the test homes under the **heterogeneous** setting.

four observations: (i) ***Appropriate prior selection:*** The utilization of appropriate priors for the activations is of significant importance in learning effective models for disaggregation. With the benefits of Mixture-of-Gammas prior, more accurate activation distributions could be captured, thus enabling the performance achieved by VM-CGS and BSC-MoG to be better than that of DDSC+SF and DDSC, respectively. (ii) ***Bayesian discriminative learning:*** Given the aggregated

data, learning discriminative features for disaggregation is of critical importance. By holding the time/home invariant features (i.e., distributions of coefficients) fixed, Bayesian discriminative learning is designed to adapt the basis functions to the aggregated consumption for enhancing disaggregation performance. Comparing the results for BSC-MoG and DDSC regarding Avg. F-measure, Accuracy, and NDE validated that Bayesian discriminative learning model can achieve better results than the conventional discriminative model. (iii) ***Efficient inference process:*** When processing large datasets, such as the simulated and real-world datasets used in the experiments, Gibbs sampling will incur high computational costs, and speeding up model learning process is a priority. By utilizing the sparse structures of the coefficients, CGS is developed to accelerate the inference process. The comparisons between VM-CGS and VM-GS confirmed that CGS could draw equivalent samples to those of conventional Gibbs sampling but require substantially fewer operations.

7 CONCLUSION

This article presents a sparse coding-based statistical framework for low-sampling-rate water disaggregation. By applying Mixture-of-Gammas as the prior distribution of coefficients, both sparseness and non-negativeness can be inherently guaranteed, and the distribution of both active and nonactive coefficients can also be captured. Bayesian modeling of the discriminative structure is a great enhancement of disaggregation performance. The CGS method developed for the fast learning of VM has been experimentally validated to be both effective and efficient. Using large-scale synthetic and real datasets, our experimental results showed that VM significantly outperformed baselines at both the whole-home and device levels.

APPENDIXES

A SHOWCASE OF DISAGGREGATION VIA SPARSE CODING

The disaggregation process estimates device-level consumption from aggregated consumption. Figure 13 shows an example of real data observation sequences, where devices' true consumption curves are the perfect disaggregation results. With the fixture-level consumption data, we can use sparse coding to learn the dictionaries for each device, which can achieve minimal reconstruction errors for the corresponding device. For example, the dictionary of Faucet can reconstruct the consumption of Faucet better than the dictionaries learned for other devices. By combining the learning dictionaries, we can optimize the objective function of sparse coding to attain the coefficients for each device. Then, the consumption of devices can be estimated by multiplying the devices' dictionary with the corresponding coefficients. The ultimate goal is to make devices' estimated consumption as close to the true consumption as possible.

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 The Generation of a Representative Synthetic Dataset

The goal here is to design a data generator to produce a representative synthetic dataset with a sampling rate of 1/900 Hz. The generator consists of three components: event dictionary construction, frequency pattern learning, and data generation.

(1) ***Event dictionary construction*****:** Five water events (corresponding to five water devices) are considered: Faucet, Dishwasher, Toilet, Shower, and Clothes Washer. The event dictionary was built based on the real-world dataset, where each event type pointed to a set of event records for this particular event type.

(2) ***Learning frequency patterns:*** The daily and interval frequency of events was statistically calculated, where daily frequency was the number of events that happen in one day while

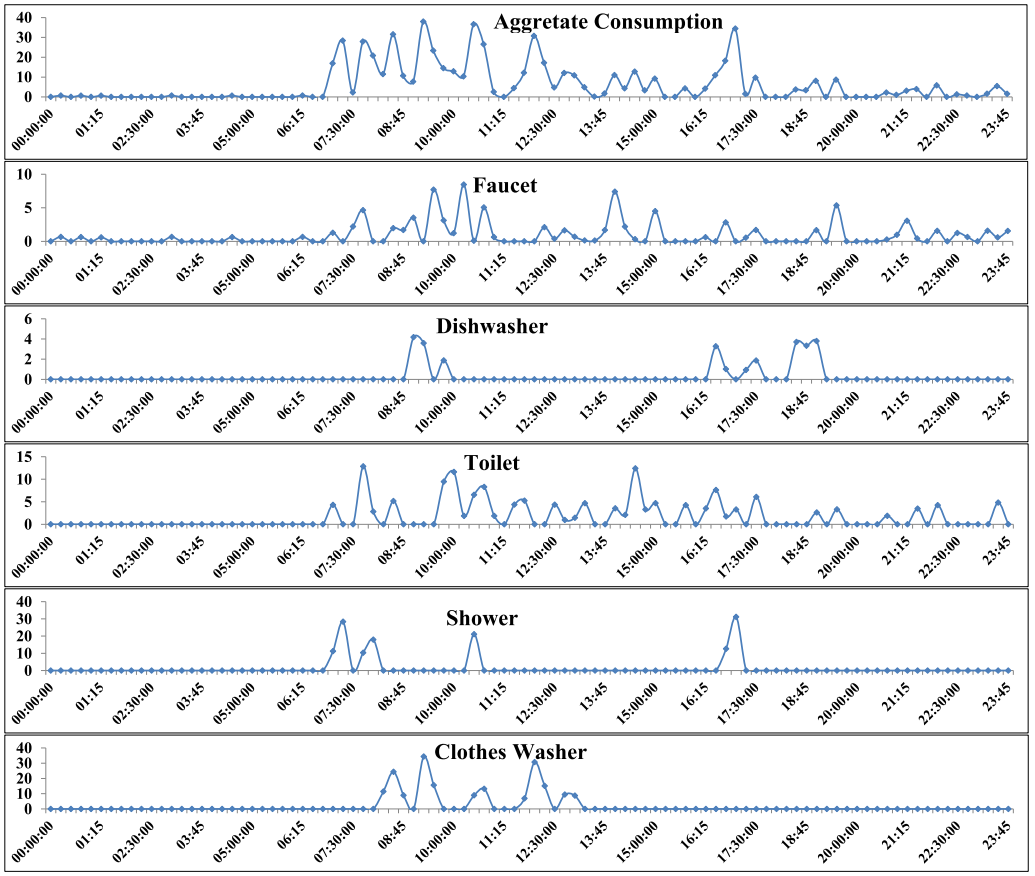

Fig. 13. Example real data observation sequences for the water disaggregation experiments.

Table 4. Statistically Estimated Parameters for Poisson Distribution, Where λ Was the Expected Value of Daily Frequency

Parameter \ Event Type	Faucet	Dishwasher	Toilet	Shower	Clothes Washer
λ	42.0856	1.0784	12.9203	2.3668	2.1761

interval frequency was the number of events starting from the corresponding interval. Based on the assumption that daily frequency of events followed a Poisson distribution, the data were used to fit a Poisson distribution with a maximum-likelihood estimation, and the estimated parameters of all events are shown in Table 4. The Cumulative Distribution Function (CDF) of the interval frequency was estimated with a kernel density function, and the PDFs and CDFs of events are respectively illustrated in Figure 14 and Figure 15.

(3) ***Data generation:*** The data were generated day by day. For each day, the Poisson distribution, which was trained with daily frequency, was first used to sample the number of events for one day. Then with the CDFs learned with the starting interval frequency, the starting intervals of the events were sampled for a particular day. Finally, the event records were randomly selected from the event dictionary. Using this procedure, the simulation data were generated for 50, 100, 400, 800, and 1,000 days to perform scalable evaluations.

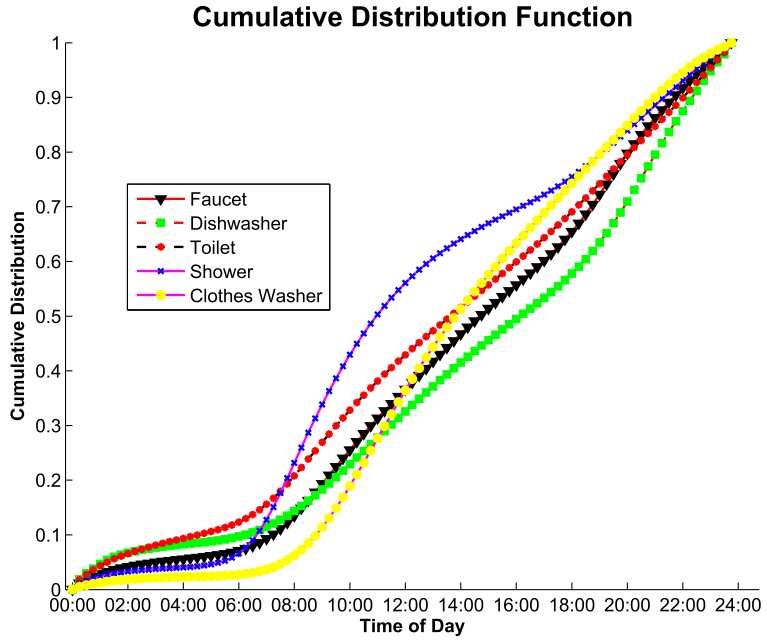

Fig. 14. CDFs of events' starting interval.

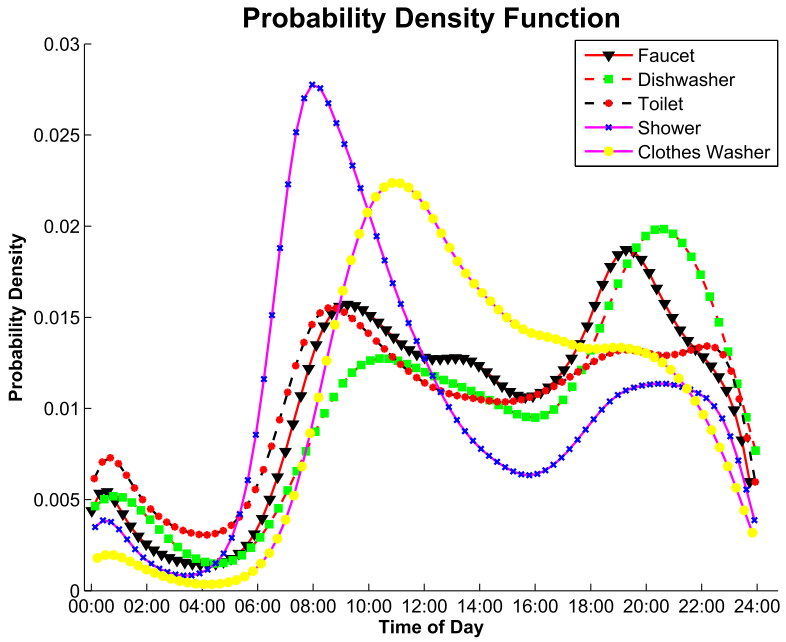

Fig. 15. PDFs of events' starting interval.

Figure 14 shows the CDFs of the starting interval for the five devices, while the corresponding Probability Density Functions (PDFs) are shown in Figure 15. Checking the CDF and PDF of Faucet, we speculate that people use more Faucets before/after breakfast (07:00–09:00) or before/after dinner (17:30–20:00). For Dishwasher, we observe that it is used more frequently in the evening (18:00–20:00), and indicates that people like to wash dishes after dinner. With respect to Toilet, as expected, more Toilets are used before/after getting up (07:00–09:00). The patterns of Shower are the most distinctive: People take a Shower in the morning (06:00–08:00) or evening (19:00–21:00). Based on the observation of Clothes Washer, we find that morning (but not that obviously) is preferred by people for washing clothes.

B.2 Prior versus Posterior over Heterogeneous Evaluations

The results of prior versus posterior for devices Faucet, Dishwasher, and Toilet are respectively shown in Figures 16, 17, and 18. All of them illustrate that the Mixture-of-Gamma prior is capable of capturing the large peak for the nonactive coefficients and the small peak for the active coefficients.

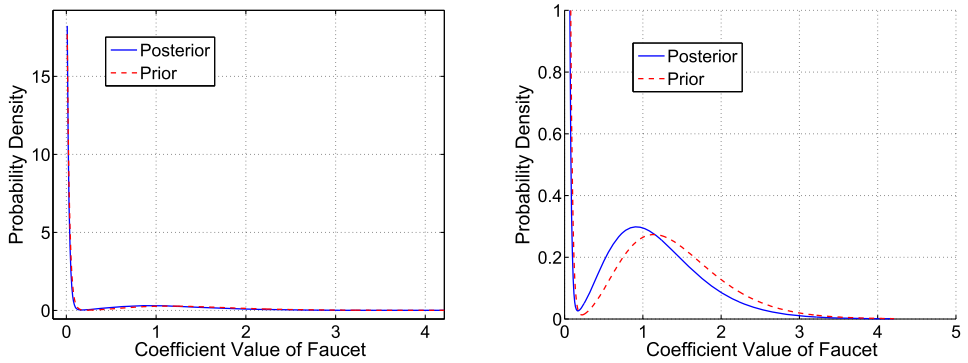

Fig. 16. A mixture of two Gammas prior and smoothed posterior of Faucet's coefficients. **Left:** Whole view of prior versus posterior. Both prior and posterior have two peaks: A large peak near zero indicates the sparse structure of the activations, and the small peak near 1 indicates a cluster of active coefficients. **Right:** Close-up of the area near the small peak showing the comparison of the prior and posterior.

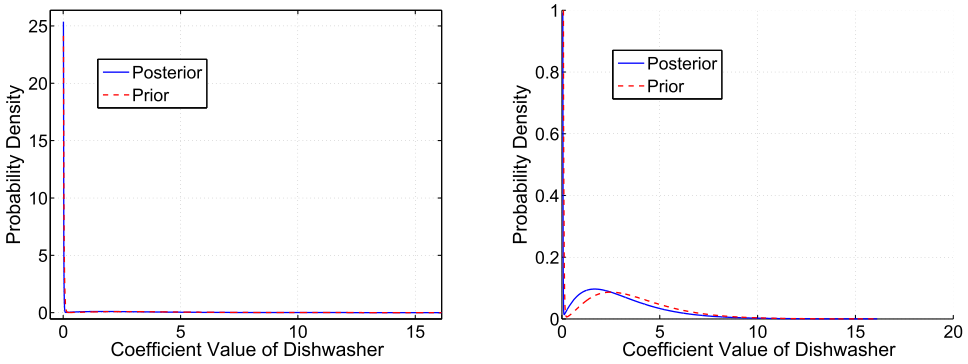

Fig. 17. A mixture of two Gammas prior and smoothed posterior of Dishwasher's coefficients. **Left:** Whole view of prior versus posterior. Both prior and posterior have two peaks: A large peak near zero indicates the sparse structure of the activations, and the small peak near 1.5 indicates a cluster of active coefficients. **Right:** Close-up of the area near the small peak showing the comparison of the prior and posterior.

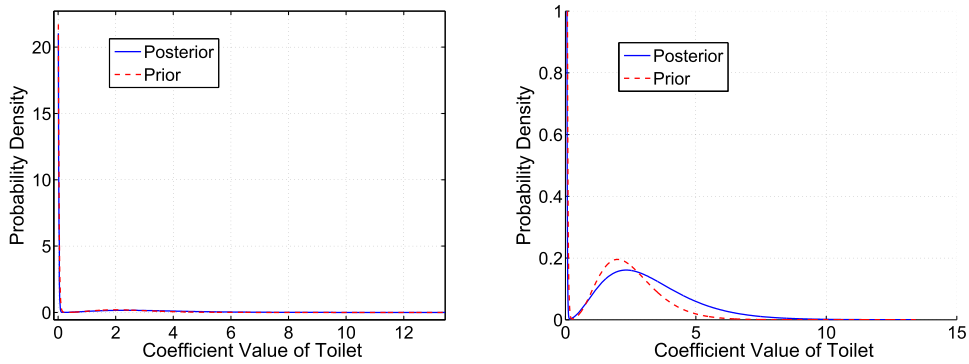

Fig. 18. A mixture of two Gammas prior and smoothed posterior of Toilet's coefficients. **Left:** Whole view of prior versus posterior. Both prior and posterior have two peaks: A large peak near zero indicates the sparse structure of the activations, and the small peak near 2 indicates a cluster of active coefficients. **Right:** Close-up of the area near the small peak showing the comparison of the prior and posterior.

ACKNOWLEDGMENT

This material is based on work supported in part by the U.S. Army Research Office under contract number W911NF-12-1-0445.

REFERENCES

H. Attias. 1999. Independent factor analysis. *Neural Computation* 117 (1999), 803–851.

F. Chen, J. Dai, B. Wang, S. Sahu, M. Naphade, and C.-T. Lu. 2011. Activity analysis based on low sample rate smart meters. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, California, USA, ACM New York, NY, 240–248.

S. S. Chen, D. L. Donoho, and M. A. Saunders. 1998. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* 20, 1 (1998), 33–61.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39, 1 (1977), 1–38.

H. Dong, B. Wang, and C.-T. Lu. 2013. Deep sparse coding based recursive disaggregation model for water conservation. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 2804–2810.

D. L. Donoho. 2006. Compressed sensing. *IEEE Transactions on Information Theory* 52, 4 (2006), 1289–1306.

B. Ellert, S. Makonin, and F. Popowich. 2016. Appliance water disaggregation via non-intrusive load monitoring (NILM). In *Smart City 360°. SmartCity 360 2016, SmartCity 360 2015*, A. Leon-Garcia et al. (Eds). Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 166. Springer, 455–467.

K. Engan, S. O. Aase, and H. J. Hakon.1999. Method of optimal directions for frame design. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5. 2443–2446.

J. Froehlich, E. Larson, T. Campbell, C. Haggerty, J. Fogarty, and S. N. Patel. 2009. HydroSense: Infrastructure-mediated single-point sensing of whole-home water activity. In *Proceedings of the International Conference on Ubiquitous Computing*. 235–244.

J. Froehlich, E. Larson, E. Saba, T. Campbell, L. Atlas, J. Fogarty, and S. Patel. 2011. A longitudinal study of pressure sensing to infer real-world water usage events in the home. In *Proceedings of the 9th International Conference on Pervasive Computing*. 50–69.

J. E. Froehlich, L. Findlater, M. Ostergren, S. Ramanathan, J. Peterson, I. Wragg, E. Larson, F. Fu, M. Bai, S. N. Patel, and J. A. Landay. 2012. The design and evaluation of prototype eco-feedback displays for fixture-level water usage data. In *Proceedings of the ACM Annual Conference on Human Factors in Computing Systems*. 2367–2376.

P. Garrigues and B. A. Olshausen. 2010. Group sparse coding with a laplacian scale mixture prior[C]. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS'10)*. Vol. 1. Vancouver, British Columbia, Canada, Curran Associates Inc., USA, 676–684.

S. Geman and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1984), 721–741.

R. Gerwen, S. Jaarsma, and R. Wilhite. 2006. Smart metering. *Leonardo-Energy.org* (2006).

M. Gharavi-Alkhansari and T. S. Huang. 1998. A fast orthogonal matching pursuit algorithm. In *Proceeddings of the IEEE International Conference on Acoustics, Speech and Signal*, Vol. 3. 1389–1392.

N. Gilbert. 2010. Balancing water supply and wildlife. *Nature News doi:10.1038/news.2010.505* (2010).

W. R. Gilks, N. G. Best, and K. K. C. Tan. 1995. Adaptive rejection metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society* 44, 4 (1995), 455–472.

P. H. Gleick. 2000. Water futures: A review of global water resources projections. *World Water Scenarios: Analyses* (2000), 27–45.

I. F. Gorodnitsky and B. D. Rao. 1997. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing* 45, 3 (1997), 600–616.

R. Grosse, R. Raina, H. Kwong, and A. Ng. 2007. Shift-invariance sparse coding for audio classification. In *Proceedings of the 23rd Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*. 149–158.

P. O. Hoyer. 2002. Non-negative sparse coding. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*. 557–565.

Anders Huss. 2012. Hybrid model approach to appliance load disaggregation: Expressive appliance modelling by combining convolutional neural networks and hidden semi-Markov models. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*. AAAI Press, Toronto, Ontario, Canada, 356–362.

Z. Jiang, Z. Lin, and L. S. Davis. 2011. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1697–1704.

Jack Kelly and William Knottenbelt. 2015. Neural nilm: Deep neural networks applied to energy disaggregation. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM, 55–64.

H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han. 2011. Unsupervised disaggregation of low frequency power measurements. In *Proceedings of the 2011 SIAM International Conference on Data Mining*. 747–758.

J. Z. Kolter, S. Batra, and A. Y. Ng. 2010. Energy disaggregation via discriminative sparse coding. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS'10)*. Vol. 1. Vancouver, British Columbia, Canada, Curran Associates Inc., USA, 1153–1161.

J. Zico Kolter and T. Jaakkola. 2012. Approximate inference in additive factorial HMMs with application to energy disaggregation. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research*, Vol. 22. La Palma, Canary Islands, PMLR, 1472–1482.

K. P. Körding, C. Kayser, and P. König. 2003. On the choice of a sparse prior. *Reviews in the Neurosciences* 14, 1–2 (2003), 53–62.

K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. Lee, and T. J. Sejnowski. 2003. Dictionary learning algorithms for sparse representation. *Neural Computation* 15, 2 (2003), 349–396.

E. Larson, J. Froehlich, T. Campbell, C. Haggerty, L. Atlas, J. Fogarty, and S. N. Patel. 2012. Disaggregated water sensing from a single, pressure-based sensor: An extended analysis of HydroSense using staged experiments. *Pervasive and Mobile Computing* 8 (2012), 82–102.

M. S. Lewicki and T. J. Sejnowski. 2000. Learning overcomplete representations. *Neural Computation* 12, 2 (2000), 337–365.

Ye Liu, Yu Zheng, Yuxuan Liang, Shuming Liu, and David S. Rosenblum. 2016. Urban water quality prediction based on multi-task multi-view learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, New York, USA, 2576–2582.

J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach. 2009. Supervised dictionary learning. In *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc., 1033–1040.

Stephen Makonin, Fred Popowich, Ivan V Bajić, Bob Gill, and Lyn Bartram. 2016. Exploiting hmm sparsity to perform online real-time nonintrusive load monitoring. *IEEE Transactions on Smart Grid* 7, 6 (2016), 2575–2585.

S. G. Mallat and Z. Zhang. 1993. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* 41, 12 (1993), 3397–3415.

Lukas Mauch and Bin Yang. 2016. A novel DNN-HMM-based approach for extracting single loads from aggregate power signals. In *Proceedings of the IEEE 2016 Acoustics, Speech and Signal Processing International Conference (ICASSP)*. IEEE, 2384–2388.

P. Mayer, W. DeOreo, E. M. Opitz, J. C. Kiefer, W. Y. Davis, B. Dziegielewski, and J. O. Nelson. 1999. *Residential End Uses of Water*. AWWA Research Foundation and American Water Works Association.

K. A. Nguyen, R. A. Stewart, and H. Zhang. 2013a. An intelligent pattern recognition model to automate the categorisation of residential water end-use events. *Environmental Modelling & Software* 47, C (2013), 108–127.

K. A. Nguyen, H. Zhang, and R. A. Stewart. 2013b. Development of an intelligent model to categorise residential water end use events. *Journal of Hydro-Environment Research* 7, 3 (2013), 182–201.

B. A. Olshausen and D. J. Field. 1996. Natural image statistics and efficient coding. *Network: Computation in Neural Systems* 7, 2 (1996), 333–339.

B. A. Olshausen and D. J. Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* 37 (1997), 3311–3325.

B. A. Olshausen and K. J. Millman. 2000. Learning sparse codes with a mixture-of-gaussians prior. In *Advances in Neural Information Processing Systems*. 841–847.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. 1993. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the IEEE Conference on Signals, Systems and Computers*. 40–44.

S. T. Roweis. 2001. One microphone source separation. In *Advances in Neural Information Processing Systems 13*. MIT Press, 793–799.

M. Schmidt, J. Larsen, and F.-T. Hsiao. 2007. Wind noise reduction using non-negative sparse coding. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*. 431–436.

M. Schmidt and R. Olsson. 2006. Single-channel speech separation using sparse non-negative matrix factorization. In *International Conference on Spoken Language Processing*.

Shikha Singh, Manoj Gulati, and Angshul Majumdar. 2016. Greedy deep disaggregating sparse coding[C]. In *Proceeding of the 3rd International Workshop on Non-Intrusive Load Monitoring (NILM'16)*.

V. Srinivasan, J. Stankovic, and K. Whitehouse. 2011. Watersense: Water flow disaggregation using motion sensors. In *Proceedings of the 3rd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. 19–24.

A. Vickers. 2001. *Handbook of Water Use and Conservation: Homes, Landscapes, Industries, Businesses, Farms*. WaterPlow Press.

T. Virtanen. 2004. Separation of sound sources by convolutive sparse coding. In *Proceedings of the Workshop on Statistical and Perceptual Audio Processing (SAPA)*.

B. Wang, F. Chen, H. Dong, A. P. Boedihardjo, and C.-T. Lu. 2012. Signal disaggregation via sparse coding with featured discriminative dictionary. In *Proceedings of the IEEE International Conference on Data Mining*. 1134–1139.

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. 2009. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 2 (2009), 210–227.

T. T. Wu and K. Lange. 2008. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* (2008), 224–244.

Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel Goddard, and Charles Sutton. 2016. Sequence-to-point learning with neural networks for nonintrusive load monitoring. *arXiv preprint arXiv:1612.09106* (2016).

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 38.

Received December 2016; revised September 2017; accepted September 2017