Online Gaussian LDA for Unsupervised Pattern Mining from Utility Usage Data

Saad Mohamad
Department of Computing
Bournemouth University
Poole, UK
saad.mohamad@outlook.com

Abdelhamid Bouchachia
Department of Computing
Bournemouth University
Poole, UK
abouchachia@bournemouth.ac.uk

Abstract—Non-intrusive load monitoring (NILM) aims at separating a whole-home energy signal into its appliance components. Such method can be harnessed to provide various services to better manage and control energy consumption (optimal planning and saving). NILM has been traditionally approached from signal processing and electrical engineering perspectives. Recently, machine learning has started to play an important role in NILM. While most work has focused on supervised algorithms, unsupervised approaches can be more interesting and of practical use in real case scenarios. Specifically, they do not require labelled training data to be acquired from individual appliances and the algorithm can be deployed to operate on the measured aggregate data directly. In this paper, we propose a fully unsupervised NILM framework based on Bayesian hierarchical mixture models. In particular, we develop a new method based on Gaussian Latent Dirichlet Allocation (GLDA) in order to extract global components that summarise the energy signal. These components provide a representation of the consumption patterns. Designed to cope with big data, our algorithm, unlike existing NILM ones, does not focus on appliance recognition. To handle this massive data, GLDA works online. Another novelty of this work compared to the existing NILM is that the data involves different utilities (e.g, electricity, water and gas) as well as some sensors measurements. Finally, we propose different evaluation methods to analyse the results which show that our algorithm finds useful patterns.

Index Terms—Unsupervised non-intrusive load monitoring, pattern recognition, Bayesian hierarchical mixture model, online Gaussian LDA.

I. Introduction

The monitoring of human behaviour is highly relevant to many real-word domains such as safety, security, health and energy management. Research on human activity recognition (HAR) has been the key ingredient to extract pattern of human behaviour. There are three main types of HAR, sensor-based [1], vision-based [2] and radio-based [3]. A common feature of these methods is that they all require equipping the living environment with embedded devices (sensors). On the other hand, non-intrusive load monitoring (NILM) requires only single meter per house or a building that measures aggregated electrical signals at the entry point of the meter. Various techniques can then be used to disaggregate per-load power consumption from this composite signal providing energy consumption data at an appliance level granularity. In this

sense, NILM focus is not extracting general human behaviour patterns but rather identifying the appliances in use. This, however, can provide insight into the energy consumption behaviour of the residents and therefore can express users' life style in their household. The idea of abandoning the high costs induced by various sensors entailed by traditional HAR makes NILM an attractive approach to exploit in general pattern recognition problems. On the other hand, taking the human behaviour into account can leverage the performance of NILM; thus, providing finer understanding of the resident's energy consumption behaviour. In this paper, we do not distinguish between patterns and appliances recognition. The main goal of our approach is to encode the regularities in a massive amount of energy consumption data into a reduced dimensionality representation. This is only possible by the fact that human behaves in certain pattern and not randomly. We are also lucky to have an extra large amount of real world data which makes this approach more viable.

Since the earliest work on NILM [4], most NILM work has been based on signal processing and engineering approaches [5], [6]. This can explain the fact that even with the economical attractive tools that NILM can provide for PR and HAR communities, it has not been widely exploited. Most of existing machine learning approaches to NILM adopt supervised algorithms [4], [7]–[13]. Such algorithms could damage the attractiveness of NILM as they require individual appliance data for training, prior to the system deployment. Hence, there is a need to install one energy meter per appliance to record appliance-specific energy consumption. This incurs extra costs and a complex installation of sensors on every device of interest. In contrast, unsupervised algorithms can be deployed to operate directly from the measured aggregate data with no need for annotation. Hence, unsupervised algorithms are clearly more suitable for NILM. To the best of our knowledge, all existing unsupervised approaches to NILM [14] concentrate on disaggregating the whole house signal into its appliances' ones. In contrast, our approach, as mentioned earlier, does not focus on identifying per-appliance signal. We instead propose a novel approach that seeks to extract human behaviour patterns from home utility usage data. These patterns could be exploited for HAR as well as energy efficiency applications.
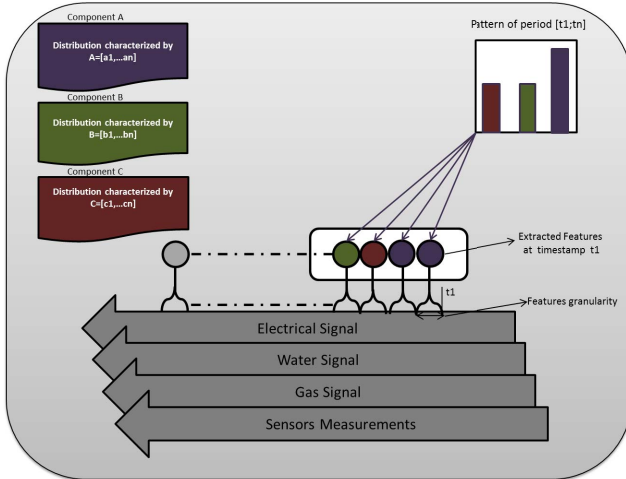
Fig. 1: Elements of the proposed approach

The proposed approach is based on a hierarchical Bayesian mixture model. More precisely, this model is a member of the family of graphical models proposed by [15] where observations, global hidden variables, local hidden variables, and fixed parameters are involved. Under further assumptions in addition to the ones in [15], we end up with a Gaussian version of Latent Dirichlet Allocation (GDLA) where the observations are continuous and not discrete. In particular, we assume that the hidden local variables are conditionally independent; hence, the observations can be treated as a bag of words. This approach has drawn inspiration from the success that LDA has achieved in the domain of text modelling. To explain the analogy between LDA and the proposed approach, we show in Fig. 1 an example where three components have been extracted from the utility usage data. Here, the components are equivalent to topics in LDA. Because the features extracted from the data are in continuous space, the components represent Gaussian distributions over the input features instead of categorical distributions over words as in LDA. A pattern is a mixture of components generating the input features over a fixed period of time. In LDA, patterns are associated with documents that can be expressed by mixture of corpus-wide topics. One can clearly notice that this bag-of-words assumption, where temporal dependency in the data is neglected, is a major simplification. However, this simplification leads to methods that are computationally efficient. Such computational efficiency is essential in our case where massive amount of data (around 4 Tb) is used to train the model.

In this work, we demonstrate that, similar to LDA in the domain of text mining, this approach can capture significant statistical structure in a specified window of data over a period of time. This structure provides understanding of regular patterns in the human behaviour that can be harnessed to provide various services including services to improve energy efficiency. For example, understanding of the usage and energy consumption patterns could be used to predict the power demand (load forecasting), to apply management policies and to avoid overloading the energy network. Moreover, providing consumers with information about their consumption behaviour and making them aware of abnormal consumption patterns compared to others can influence their behaviour to moderate energy consumption [16].

As already mentioned, this algorithm is going to be trained over a very huge amount of data resulting from the high sampling rate around 205 kHz of the electricity signal which gives us an advantage compared to the data used in other research studies except for [17]–[19]. Specifically, high sampling rate allows extraction of rich features in contrast to the limited number of features that can be extracted from low frequency data. To handle such big amount of data, online version of GLDA is developed. This can be done by defining particular distributions for the exponential family in the class of models described in [15]. More details can be found in Sec. III. Besides the advantage the data size offers, apart from [20], [21] whose sampling rate is very low, our data is the only one including water and gas usage data. Moreover, measurements provided by additional sensors are also exploited to refine the performance of the pattern recognition algorithm. More details on the data can be found in Sec. IV. The diversity of the data is another motivation for adopting a pattern recognition approach rather than traditional disaggregation approach.

The rest of the paper is organised as follows. Section II presents the related work. Section III presents the proposed approach. Section IV describes the data and discusses the obtained results. Finally, Sec. V concludes the paper and hints to future work.

II. RELATED WORK

We divide the related work into two parts: (i) machine learning approaches to NILM and (ii) NILM data used in the literature.

As we have discussed in the introduction, most of existing NILM studies are not based on machine learning algorithms and most of machine learning NILM algorithms are supervised ones [4], [7]–[13]. Such algorithms requires training on labelled data which is expensive and laborious to obtain. In fact, the practicality of NILM is stemmed from the fact that it comes with almost no setup cost. Recently, researchers have started exploring unsupervised machine learning algorithms to NILM. These methods have mainly focused on performing energy disaggregation to discern appliances from the aggregated load data directly without performing any sort of event detection. The most prominent of these methods are based on Dynamic Bayesian Network models, in particular different variants of Hidden Markov Model (HMM) [22]–[24].

Authors in [22] proposes to use Factorial Hidden Markov Model (FHMM) and three of its variants: Factorial Hidden Semi-Markov Model (FHSMM), Conditional FHMM (CFHMM) and Conditional FHSMM (CFHSMM) to achieve energy disaggregation. The main idea is that the dynamics of the state occupancy of each appliance evolves independently and the observed aggregated signal is some joint function of all the appliances states. To better model the state occu-

42

pancy duration, that is modelled with a geometric distribution by FHMM, authors propose to use FHSMM which allows modelling the durations of the appliances states with gamma distribution. Authors also propose CFHMM to incorporate additional features, such as time of day, other sensor measurements, and dependency between appliances. To harness the advantages of FHSMM and CFHMM, authors propose a combination of the two models resulting in CFHSMM. In that work, the electricity signal was sampled at low frequency which is in contrast to our work.

Similar approach was taken in [23] where Additive Factorial Hidden Markov Model (AFHMM) was used to separate appliances from the aggregated load data. The main motivation and contribution of this approach is that it addresses the local optima problems that existing approximate inference techniques [22] are highly susceptible to experience. The idea is to exploit the additive structure of AFHMM to develop a convex formulation of approximate inference that is more computationally efficient and has no issues of local optima. Although, this approach was applied on relatively high frequency electricity data [19], the data scale is not close to ours. Hierarchical Dirichlet Process Hidden Semi-Markov Model (HDP-HSMM) is used in [24] to incorporate duration distributions (Semi Markov) and allows to infer the number of states from the data (Hierarchical Dirichlet Process). On the contrary, the AFHMM algorithm in [23] requires the number of appliances (states) to be set a-priori.

The common feature of the approaches discussed so far is that the considered data sets are collected only from the electricity signals. In contrast, our data involves different utilities namely electricity, water and gas data as well some sensors measurements that provide contextual features. To the best of our knowledge, the only data that considers water and gas usage data is [20], [21]. However, the sampling rate of this data is very low compared to ours. Authors in [25] exploit the correlation between appliances and side information, in particular temperature, in a convex optimisation problem for energy disaggregation. This algorithm is applied on low sampling rate electricity data with contextual supervision in the form of temperature information.

To wrap up this section, three features distinguish our approach from existing ones. It bridges the gap between pattern recognition and NILM making it beneficial for a variety of different applications. Driven by massive amount of data, our method is computationally efficient and scalable, unlike state-of-the-art probabilistic methods that posit detailed temporal relationships and involve complex inference steps. The available data has a high sampling rate electricity data allowing extracting more informative features and includes data from other utility usage and additional sensors measurements. Thus, our work also covers the research aspect of NILM concerned with the acquisition of data, prepossessing steps and evaluation of NILM algorithms.

III. THE APPROACH

In this section, we present the proposed approach which consists of two steps: features extraction and pattern mining.

First, a background on stochastic variational inference for a family of graphical models is provided. Next, we derive the pattern mining algorithm, online GLDA, which is an instance of the family of graphical models and operates online to accommodate high volume and speed data streams. Finally, we present the feature extraction step and summarise the full algorithm.

A. Background: Stochastic Variational Inference

In the following, we describe the model family of which GLDA is a member and review SVI.

Model family. The family of models considered here consists of three random variables: observations $\boldsymbol{x} = \boldsymbol{x}_{1:D}$, local hidden variables $\boldsymbol{z} = \boldsymbol{z}_{1:D}$, global hidden variables $\boldsymbol{\beta}$ and fixed parameters $\boldsymbol{\alpha}$. The model assumes that the distribution of the D pairs of $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ is conditionally independent given $\boldsymbol{\beta}$. Furthermore, their distribution and the prior distribution of $\boldsymbol{\beta}$ belong to the exponential family.

$$p(\boldsymbol{\beta}, \boldsymbol{x}, \boldsymbol{z} | \boldsymbol{\alpha}) = p(\boldsymbol{\beta}|\boldsymbol{\alpha}) \prod_{i=1}^{D} p(\boldsymbol{z}_i, \boldsymbol{x}_i | \boldsymbol{\beta}) \quad (1)$$

$$p(\boldsymbol{z}_i, \boldsymbol{x}_i | \boldsymbol{\beta}) = h(\boldsymbol{x}_i, \boldsymbol{z}_i) \exp\left(\boldsymbol{\beta}^T t(\boldsymbol{x}_i, \boldsymbol{z}_i) - a(\boldsymbol{\beta})\right) \quad (2)$$

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = h(\boldsymbol{\beta}) \exp\left(\boldsymbol{\alpha}^T t(\boldsymbol{\beta}) - a(\boldsymbol{\alpha})\right) \quad (3)$$

Here, we overload the notation for the base measures $h(.)$, sufficient statistics $t(.)$ and log normalizer $a(.)$. While the soul of the proposed approach is generic, for simplicity we assume a conjugacy relationship between $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ and $\boldsymbol{\beta}$. That is, the distribution $p(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{z})$ is in the same family as the prior $p(\boldsymbol{\beta}|\boldsymbol{\alpha})$.

Note that this innocent looking family of models includes (but is not limited to) latent Dirichlet allocation [26], Bayesian Gaussian mixture, probabilistic matrix factorization, hidden Markov models, hierarchical linear and probit regression, and many Bayesian non-parametric models.

Mean-field variational inference. Variational inference (VI) approximates intractable posterior $p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{x})$ by positing a family of simple distributions $q(\boldsymbol{\beta}, \boldsymbol{z})$ and find the member of the family that is closest to the posterior (closeness is measured with KL divergence). The resulting optimization problem is equivalent maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(q) = E_q[\log p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta})] - E_q[\log p(\boldsymbol{z}\boldsymbol{\beta})] \leq \log p(\boldsymbol{x}) \quad (4)$$

Mean-field is the simplest family of distribution, where the distribution over the hidden variables factorizes as follows:

$$q(\boldsymbol{\beta}, \boldsymbol{z}) = q(\boldsymbol{\beta}|\boldsymbol{\lambda}) \prod_{i=1}^{D} p(\boldsymbol{z}_i|\boldsymbol{\phi}_i) \quad (5)$$

Further, each variational distribution is assumed to come from the same family of the true one. Mean-field variational inference optimizes the new ELBO with respect to the local and global variational parameters $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$.

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\phi}) = E_q\left[\log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right] + \sum_{i=1}^{D} E_q\left[\log \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\beta})}{q(\boldsymbol{z}_i)}\right] \quad (6)$$

43

It iteratively updates each variational parameter holding the other parameters fixed. With the assumptions taken so far, each update has a closed form solution. The local parameters are a function of the global parameters.

$$\phi(\boldsymbol{\lambda}_t) = \arg\max_{\phi} \mathcal{L}(\boldsymbol{\lambda}_t, \phi) \tag{7}$$

We are interested in the global parameters which summarise the whole dataset (clusters in Bayesian Gaussian mixture, topics in LDA).

$$\mathcal{L}(\boldsymbol{\lambda}) = \max_{\phi} \mathcal{L}(\boldsymbol{\lambda}, \phi) \tag{8}$$

To find the optimal value of $\boldsymbol{\lambda}$ given that ϕ is fixed, we compute the natural gradient of $\mathcal{L}(\boldsymbol{\lambda})$ and set it to zero by setting

$$\boldsymbol{\lambda}^* = \boldsymbol{\alpha} + \sum_{i=1}^{D} E_{\phi_i(\boldsymbol{\lambda}_t)}[t(\boldsymbol{x}_i, \boldsymbol{z}_i)] \tag{9}$$

Thus, the new optimal global parameters are $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}^*$. The algorithm works by iterating between computing the optimal local parameters given the global ones (Eq. 7) and computing the optimal global parameters given the local ones (Eq. 9).

Stochastic variational inference. Rather than analysing all the data to compute $\boldsymbol{\lambda}^*$ at each iteration, stochastic optimization can be used. Assuming that the data samples are uniformly randomly selected from the dataset, an unbiased noisy estimator of $\mathcal{L}(\boldsymbol{\lambda}, \phi)$ can be developed based on a single data point.

$$\mathcal{L}_i(\boldsymbol{\lambda}, \phi_i) = E_q\left[\log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right] + D E_q\left[\log \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i | \boldsymbol{\beta})}{q(\boldsymbol{z}_i)}\right] \tag{10}$$

The unbiased stochastic approximation of the ELBO as a function of $\boldsymbol{\lambda}$ can be written as follows

$$\mathcal{L}_i(\boldsymbol{\lambda}) = \max_{\phi_i} \mathcal{L}_i(\boldsymbol{\lambda}, \phi_i) \tag{11}$$

Following the same steps in the previous section, we end up with a noisy unbiased estimate of Eq. 8

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\alpha} + D E_{\phi_i(\boldsymbol{\lambda}_t)}[t(\boldsymbol{x}_i, \boldsymbol{z}_i)] \tag{12}$$

At each iteration, we move the global parameters a step-size ρ_t (learning rate) in the direction of the noisy natural gradient.

$$\boldsymbol{\lambda}_{t+1} = (1 - \rho_t)\boldsymbol{\lambda}_t + \rho_t \hat{\boldsymbol{\lambda}} \tag{13}$$

With certain conditions on ρ_t, the algorithm converges ($\sum_{t=1}^{\infty} \rho_t = \infty$, $\sum_{t=1}^{\infty} \rho_t^2 < \infty$) [27].

B. Online Gaussian LDA

Gaussian LDA (GLDA), as its name suggests, is an LDA with Gaussian components over the observations in place of the multinational ones of LDA. Hence, it is an instance of the family of models described in Sec III where the global, local, observed variables and their distributions are set as follows:

- the global variables $\{\boldsymbol{\beta}\}_{k=1}^{K} \equiv \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}_{k=1}^{K}$ are the components in GLDA. A component is a distribution over the input in the feature space, where the probability of an input vector $\boldsymbol{x}$ in component k,

$p(\boldsymbol{x}|\boldsymbol{\beta}, k) = N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Hence, the prior distribution of $\boldsymbol{\beta}_k$ is a Normal-Inverse-Wishart distribution $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_k NIW(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | \boldsymbol{m}, \boldsymbol{\omega}, s, v)$.
- The local variables are the component proportions $\{\boldsymbol{\theta}_d\}_{d=1}^{D}$ and the component assignments $\{\{z_{d,i}\}_{d=1}^{D}\}_{i=1}^{n}$ which index the Gaussian component that generate the observations. Each pattern is associated with a component proportion which is a distribution over components, $p(\boldsymbol{\theta}) = \prod_d Dir(\boldsymbol{\theta}_d; \boldsymbol{\alpha})$. The assignments $\{\{z_{d,i}\}_{d=1}^{D}\}_{i=1}^{n}$ are indices, generated by $\boldsymbol{\theta}_d$, that couple components with observations, $p(\boldsymbol{z}_d | \boldsymbol{\theta}) = \prod_i \theta_{d,z_{d,i}}$.
- The observations $\boldsymbol{x}_d$ are the observations during a specified period of time which are assumed to be drawn from components $\boldsymbol{\beta}$ selected by indices $\boldsymbol{z}_d$, $p(\boldsymbol{x}_d | \boldsymbol{z}_d, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_i N(\boldsymbol{x}_{d,i} | \boldsymbol{\mu}_{z_{d,i}}, \boldsymbol{\Sigma}_{z_{d,i}})$.

The basic idea of GLDA is that each pattern is represented as random mixture over latent components, where each component is characterised by a distribution over the input observations. GLDA assumes the following generative process:

1 Draw components as follows: co-variance $\boldsymbol{\Sigma}_k \sim W^{-1}(\boldsymbol{\omega}, v)$; mean $\boldsymbol{\mu}_k \sim N(\boldsymbol{m}, \frac{1}{s}\boldsymbol{\Sigma})$ for $k \in \{1, ..., K\}$
2 Draw component proportions $\boldsymbol{\theta}_d \sim Dir(\alpha, ..., \alpha)$ for $d \in \{1, ..., D\}$

   2.1 Draw component assignments $z_{d,i} \sim Mult(\boldsymbol{\theta}_d)$ for $i \in \{1, ..., n\}$

   2.1.1 Draw an observation $x_{d,i} \sim N(\boldsymbol{\mu}_{z_{d,i}}, \boldsymbol{\Sigma}_{z_{d,i}})$

According to Sec. III-A, each variational distribution is assumed to come from the same family of the true one. Hence, $q(\boldsymbol{\beta}_k | \boldsymbol{\lambda}_k) = NIW(\boldsymbol{qm}_k, \boldsymbol{q\omega}_k, qs_k, qv_k)$, $q(\boldsymbol{\theta}_d | \boldsymbol{\gamma}_d) = Dir(\boldsymbol{\gamma}_d)$ and $q(z_{d,i} | \boldsymbol{\phi}_{d,i}) = Mult(\boldsymbol{\phi}_{d,i})$. To compute the update equations (shown in Eq. 13) of the global variational parameters for GLDA, we need to find the sufficient statistic $t(.)$ presented in Eq. 2. By writing the likelihood of GLDA in the form of Eq. 2, we can obtain the following update:

$$qs_k^{t+1} = (1 - \rho_t)qs_k^t + \rho_t(s + D\sum_{i=1}^{n} \phi_{d,i}^k)$$

$$qv_k^{t+1} = (1 - \rho_t)qv_k^t + \rho_t(v + D\sum_{i=1}^{n} \phi_{d,i}^k)$$

$$\boldsymbol{qm}_k^{t+1} = (1 - \rho_t)\frac{qs_k^t}{qs_k^{t+1}}\boldsymbol{qm}_k^t + \frac{\rho_t}{qs_k^{t+1}}(s\boldsymbol{m} + D\sum_{i=1}^{n} \phi_{d,i}^k \boldsymbol{x}_{d,i})$$

$$\boldsymbol{q\omega}_k^{t+1} = (1 - \rho_t)(\boldsymbol{q\omega}_k^t + qs_k^t \boldsymbol{qm}_k^t \boldsymbol{qm}_k^{tT}) - qs_k^{t+1}\boldsymbol{qm}_k^{t+1}$$

$$\boldsymbol{qm}_k^{t+1T} + \rho_t(\boldsymbol{\omega} + s\boldsymbol{mm}^T + D\sum_{i=1}^{n} \phi_{d,i}^k \boldsymbol{x}_{d,i}\boldsymbol{x}_{d,i}^T)$$

$$\tag{14}$$

where $d \sim \{1,...D\}$. The local variational parameters can be computed as follows:

$$\phi_{d,i}^k \propto exp\bigg(\Psi(\gamma_{d,i}) - \frac{qv_k}{2}(\boldsymbol{x}_{d,i} - \boldsymbol{qm}_k)^T \boldsymbol{q\omega}_k^{-1}(\boldsymbol{x}_{d,i} - \boldsymbol{qm}_k)$$

$$- \frac{F}{2qs_k} + \frac{1}{2}\Big(\sum_{j=1}^F \Psi(\frac{qv_k + 1 - j}{2}) + \log|\boldsymbol{q\omega}_k^{-1}| - F\log\pi\Big)\bigg)$$

$$\boldsymbol{\gamma}_d = \alpha + \sum_{i=1}^n \phi_{d,i}^k \qquad (15)$$

where F is the dimension of the feature space. Details on how Eq. 14 and 15 are derived can be found in [15], [28].

Having computed the main elements of the pattern recognition algorithm, we move to the next section in which the features extraction together with the pattern mining process are described.

C. Features Extraction and Pattern Mining

The ultimate goal of the proposed approach is to provide a lower-dimensional representation expressing patterns in the data. To this end, the proposed approach consists of two steps: (1) features extraction and (2) pattern mining. In fact, features extraction is required to avoid the effects of the curse of dimensionality when applying the pattern mining algorithm (online GLDA). It helps reduce unnecessary redundancies in the raw data signal and extract informative features. Although, appliances identification is not the goal of this work, distinctive features providing useful information to discern appliances under use will be informative for GLDA. Such information suggests the activity performed by the residents leading to insight on their behaviour. In this work, features are extracted from the electricity signal only. Data coming from the other utilities and the sensors measurements are sampled at much lower rate, hence it is of small size. It feeds GLDA in their original representation.

Two main types of features have been proposed in the literature [6] with the purpose of detecting events: steady-state and transient event-based features. Steady state methods relate to changing operation state of the appliance; for example a change of steady-state active power measurement from a high to low value can identify whether the appliance is being turned On or Off. This kind of features can be captured with low sampling rate. The transient methods capture transient behaviour between steady-states; for example high frequency noise in electrical current or voltage, as a result of an appliance changing operation state. This type of features requires high sampling rate. Examples of features that can define appliance state transitions are shape, size, duration and the harmonics of the transient waveforms. These two types of features have been often used with supervised machine learning for appliances identification.

On the other hand, unsupervised algorithms like the ones discussed in Sec. II work directly on separating the power signal of individual appliance from the aggregated signal without preforming any sort of event detection. In contrast, our approach comprises the feature extraction phase, however, it is not an event detection method. We extract features that harness the high frequency of the sampling by exploiting information in the frequency spectrum as well as the conventional NILM features like the well-known reactive and active power features. Real and reactive power features have been shown to be very useful (alone or accompanied with other features) in many conventional non-intrusive load monitoring approaches [4], [29]–[31]. The importance of these features or features derived from them is that they convey information about the load of the appliance as well as the nature of it (difference between reactive and active power). Another advantage is that they do not require high sampling rate and therefore expensive current and voltage meters. However, the provided electricity data is sampled at high rate.

In addition, to exploit the information offered with the high sampling rate, we extract frequency domain features. We compute the RMS spectrum power over fixed bands of frequencies. The size and number of these bands are provided as parameter of the electricity extraction function. RMS spectrum power provides information about the waveforms. Different waveforms can characterise different types of appliances. Hence, it is expected that the frequency domain features will be very useful in the mining task. Adopting these features was also inspired by the work of the researchers at MIT as well as other research studies [9], [11], [32]–[34]. The harmonics of the signal can also uniquely characterise non-linear loads that draw non-sinusoidal current during the operation. They have been used in combination with real and reactive power features [9], [35]. Detailed experiments for the features extraction are carried out in the next section.

These features are computed over windows of given duration granularity. Together with the gas water and other sensors data, they form a vector of observations. Specified number of these vectors are stacked over a pattern window to be used by the proposed algorithm (see fig. 1). Algorithm 1 summarises the steps of the proposed method.

IV. EMPIRICAL EVALUATION

The experimental data that GLDA will be tested on along with details about the data pre-processing stages are presented in App. A.

In all experiments, we use the empirical Bayes method to online point estimate the hyper-parameters from the data. The idea is to maximise the log likelihood of the data with respect to the hyper-parameters. Since the computation of the log likelihood of the data is not tractable, approximation based on the variational inference algorithm used in Sec. III-B is employed. Following the same steps used to derive Eq. 14 and Eq. 15, the update function for the hyper-parameters can be derived. The number of components is fixed to $K = 50$ We evaluated a range of settings of the learning parameters: κ, τ_0 and batch size BS on a validation set, where the parameters κ and τ_0, defined in [36], control the learning step-size ρ_t. We used the data collected during the last week for validation and testing.

In order to evaluate GLDA, we use the perplexity measure. Perplexity is used to quantify the fit of the model to the data.

45

Algorithm 1 Pattern mining for energy consumption behaviour

1: **Input:** raw-data window length, R; preprocessed-data window length, n; number of preprocessed-data windows, D; number of GLDA's components, K; total number of iterations, T; learning rate, $\{\rho_t\}_{t=0}^{T}$; GLDA's hyper-parameters, $(\boldsymbol{m},\boldsymbol{\omega},s,v)$, α.

2: **Initialisation:** variational parameters: $\{(\boldsymbol{qm}_k^0, \boldsymbol{q\omega}_k^0, qs_k^0, qv_k^0)\}_{k=0}^{K}$.

3: **for** $t = 0, 1, 2, ...T - 1$ **do**

4:    Read sequentially n raw data windows of length R.

5:    Extract features (see Sec. III-C) for each window

6:    Form windows of data points (actual input) of length n in the new feature space ($\{\boldsymbol{x}_{d,i}\}_{i=1}^{n}$).

7:    Initialise $\{\gamma_{dk}\}_{k=1}^{K}$

8:    **repeat**

9:      Compute local variational parameters $\{\{\phi_{d,i}^k\}_{i=1}^{n}\}_{k=1}^{K}$ (see Eq. 15)

10:      Update local variational parameters $\{\gamma_{dk}\}_{k=1}^{K}$ (see Eq. 15)

11:    **until** local parameters converge

12:    Update global variational parameters $\{(\boldsymbol{qm}_k^0, \boldsymbol{q\omega}_k^0, qs_k^0, qv_k^0)\}_{k=0}^{K}$ (see Eq. 14)

13: **end for**

It is defined as the reciprocal geometric mean of the inverse marginal probability of the input in the held-out test set. Since perplexity cannot be computed directly, a lower bound on it is derived in a similar way to the one in [26]. This bound is used as a proxy for the perplexity.

Moreover, to investigate the quality of the results, we study the regularity of the mined patterns by matching them across similar periods of time. For instance, it is expected that similar patterns will emerge in specific hours like breakfast in every morning, watching TV in the evening, etc. Hence, it is interesting to understand how such patterns occur as regular events.

Finally, to provide quantitative evaluation of the algorithm, we propose a mapping method that reveals the specific energy consumed for each pattern. By doing so, we can evaluate numerically the coherence of the extracted patterns by fitting a regression model to the energy consumption over components:

$$A\boldsymbol{w} = \boldsymbol{b} \qquad (16)$$

where $\boldsymbol{w}$ is a vector expressing energy consumption associated with components. $\boldsymbol{b}$ is a vector representing per-pattern consumption and A is the matrix of the per-pattern components proportions obtained by GLDA. This technique will also allow numerically checking the predicted consumption against the real consumption.

1) A- Model Fitness:

Although online GLDA converges for any valid κ, τ_0 and BS, the quality and speed of the convergence may depend on how the learning parameters are set. We run online GLDA

TABLE I: Parameter settings

Batch size: BS	1	2	4	5
Learning factor: κ	0.9	0.9	0.9	0.9
Learning delay: τ_0	1024	1024	1024	1024
Perplexity	4262700	1027100	376558	332150

on the training sets for $\kappa \in \{0.1, 0.3, 0.6, 0.8, 0.9\}$, $\tau_0 \in \{1, 64, 256, 1024\}$ and $BS \in \{1, 2, 4, 5\}$. Table I summarises the best settings of each batch size along with the perplexity obtained on the test set.

The obtained results indicate that increasing the batch-size leads to better perplexity. However, the computation complexityincreases. Hence, we made a balance between model fitness and computation by setting the batch size to 4, where the best learning parameters are $\kappa = 0.9$ and $\tau = 1024$.

2) B- Pattern Regularity:

Using the optimal parameters' setting, we examine in the following the regularity of the mined patterns. To do that, we use the last two weeks of the data (from Thursday 18-05-2017 23:45:22 to Thursday 01-06-2017 23:45:22) for testing. To study the regularity of the energy consumption behaviour of the residents, we compare the mined patterns across different days of the testing period. These patterns are represented by the proportions of the different components (topics) inferred from the training data. To visualise the patterns, we plot gray-scale images showing the probability of different components with respect to the time. Black colour indicates probability of the component = 0, while white colour indicates probability = 1. Figure 2 shows 14 figures split into two rows. The first row corresponds to the week from 18-05-2017 23:45:22 to 26-05-2017 23:45:22. The second row corresponds to the week from 26-05-2017 23:45:22 to 01-06-2017 23:45:22. Each figure depicts the pattern over 24 hours. The figures of the same days from different weeks are shown on the same column.

It can be clearly seen from these figures that there is regular patterns across rows. That is, similar energy consumption patterns appear across different weeks. Moreover, consumption patterns across working days within the same week are similar. On the other hand, for a specific week, the patterns over the weekend days and Friday are quite dissimilar to the rest within and across that week. This regularity may be caused by regular user lifestyle leading to similar energy consumption behaviour within and across the weeks. Such regularity is violated in the weekend, where more random activities could take place. Note that the difference between the patterns on 29-05-2017 (Monday) and that on 22-05-2017 (Monday) may be caused by the fact that on the 29th of May there was a bank holiday in the UK. Having shown that there is some regularity in the mined patterns, it is more likely that specific energy consumption can be associated with each component. In the next section, we apply a regression method to map the patterns (e.i., components proportions) to energy consumption. Thus, the parameters of interest are the energy consumption associated with the components. By attaching an energy consumption with each component, we can help validate the coherence of the extracted patterns and do forecasting.
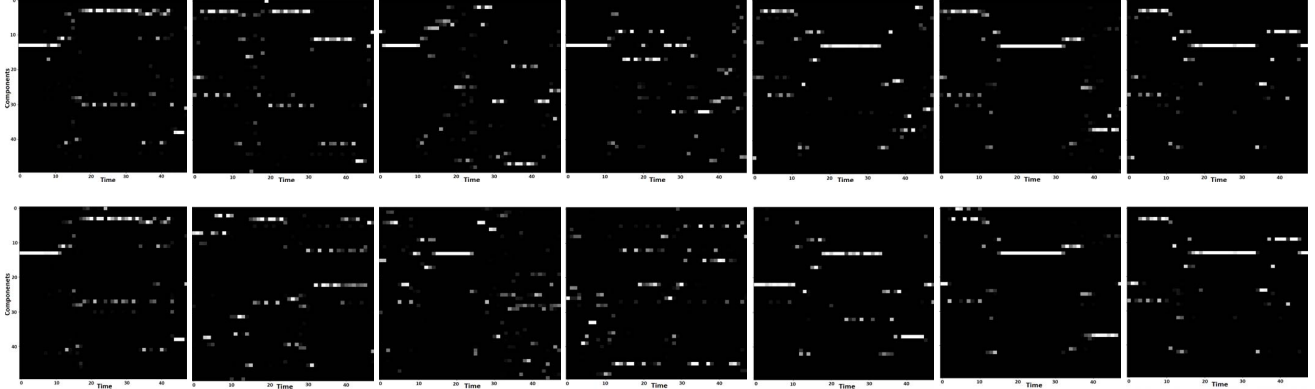
46

Fig. 2: Emergence of patterns

3) C- Energy Mapping:

As shown in the previous section, GLDA can express the energy consumption patterns by mixing global components summarising data. These global components can be thought of as a base in the space of patterns. Each component is a distribution over a high-dimensional feature space and understanding what it represents is not easy. Hence, we propose to associate electricity consumption quantities to each component. Such association is motivated by the fact that an energy consumption pattern is normally governed by the usage of different appliances in the house. There should be a strong relation between components and appliances usage. Hence, a relation between components and energy consumption is plausible. Note that the best case scenario occurs if each component is associated with the usage of a specific appliance. Apart from the coherence study, associating energy consumption with each component can be used to forecast the energy consumption. This can be done through pattern forecasting which will be investigated in future work. More details will be given in Sec. V. We apply a simple least-square regression method to map patterns to energy consumption, expressed as follows:

$$\min ||A\boldsymbol{w} - \boldsymbol{b}||^2 \qquad (17)$$

where $\boldsymbol{w}$ is the per-component energy consumption vector, $\boldsymbol{b}$ is the per-pattern consumption vector and A is the matrix of the per-pattern components' proportions which is computed by GLDA. We train the regression model on the first testing week and run the model on the second one. Figure 3 shows the energy consumption (in joules) along with the estimated consumption computed using the learned per-component consumption parameters.

The similarity between the estimated and computed energy consumption demonstrates that the LDA components express distinct usages of energy. Such distinction can be the result of the usage of different appliances likely having distinct energy consumption signatures. Thus, the proposed approach produces coherent and regular patterns that reflect the energy consumption behaviour and human activities. Note that it is possible that different patterns (or appliance usages) may have

the same energy consumption and that is why both estimated and computed energy consumption in Fig. 3 are not fully the same.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a novel approach, Gaussian LDA (GLDA), to extract patterns of the users' consumption behaviour from data involving different utilities (e.g, electricity, water and gas) as well as some sensors measurements. GLDA is fully unsupervised and works online which made it efficient for big data. To analyse the performance of GLDA, we proposed a three- step evaluation that covers: model fitness, qualitative analysis and quantitative analysis. The experiments show that the proposed method is capable of extracting regular and coherent patterns that highlight energy consumption over time.

In the future, we foresee four directions for research to improve the obtained results and provide more features: (i) developing online dynamic Gaussian latent Dirichlet allocation (DGLDA) to consider the temporal dependency in the data leading to better results and allowing forecasting, (ii) replace the engineered features with ones extracted by a deep learning model trained directly on the raw data to yield richer low-level features, (iii) develop more scalable GLDA by applying asynchronous distributed GLDA which can be derived from [37] instead of SVI and (iiii) involve active learning strategy to query users about about ambiguous or unknown activities in order to guide the learning process when needed [38], [39].

REFERENCES

[1] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[3] S. Wang and G. Zhou, "A review on radio based activity recognition," *Digital Communications and Networks*, vol. 1, no. 1, pp. 20–29, 2015.

[4] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[5] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE transactions on Consumer Electronics*, vol. 57, no. 1, 2011.

[6] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, "Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey," *Sensors*, vol. 12, no. 12, pp. 16 838–16 866, 2012.

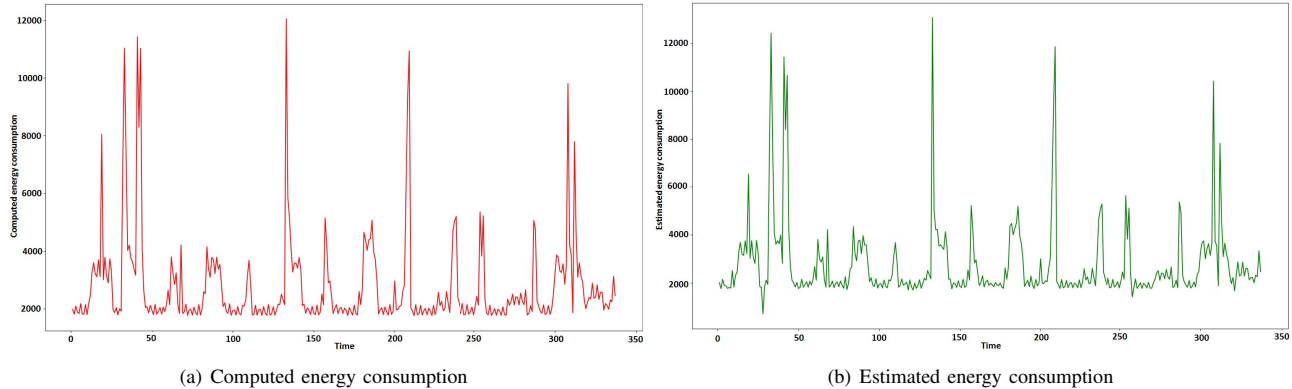(a) Computed energy consumption

(b) Estimated energy consumption

Fig. 3: Evolution of the energy consumption over time

[7] J. Liang, S. K. Ng, G. Kendall, and J. W. Cheng, "Load signature study—part i: Basic concept, structure, and methodology," *IEEE transactions on power Delivery*, vol. 25, no. 2, pp. 551–560, 2010.

[8] J. Z. Kolter, S. Batra, and A. Y. Ng, "Energy disaggregation via discriminative sparse coding," in *Advances in Neural Information Processing Systems*, 2010, pp. 1153–1161.

[9] D. Srinivasan, W. Ng, and A. Liew, "Neural-network-based signature recognition for harmonic source identification," *IEEE Transactions on Power Delivery*, vol. 21, no. 1, pp. 398–405, 2006.

[10] M. Berges, E. Goldman, H. S. Matthews, and L. Soibelman, "Learning systems for electric consumption of buildings," in *Computing in Civil Engineering (2009)*, 2009, pp. 1–10.

[11] A. G. Ruzzelli, C. Nicolas, A. Schoofs, and G. M. O'Hare, "Real-time recognition and profiling of appliances through a single electricity sensor," in *Sensor Mesh and Ad Hoc Communications and Networks (SECON), 2010 7th Annual IEEE Communications Society Conference on*. IEEE, 2010, pp. 1–9.

[12] J. Kelly and W. Knottenbelt, "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, 2015, pp. 55–64.

[13] Y.-X. Lai, C.-F. Lai, Y.-M. Huang, and H.-C. Chao, "Multi-appliance recognition system with hybrid svm/gmm classifier in ubiquitous smart home," *Information Sciences*, vol. 230, pp. 39–55, 2013.

[14] R. Bonfigli, S. Squartini, M. Fagiani, and F. Piazza, "Unsupervised algorithms for non-intrusive load monitoring: An up-to-date overview," in *Environment and Electrical Engineering (EEEIC), 2015 IEEE 15th International Conference on*. IEEE, 2015, pp. 1175–1180.

[15] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[16] C. Fischer, "Feedback on household electricity consumption: a tool for saving energy?" *Energy efficiency*, vol. 1, no. 1, pp. 79–104, 2008.

[17] J. Kelly and W. Knottenbelt, "The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes," *Scientific data*, vol. 2, p. 150007, 2015.

[18] A. Filip, "Blued: A fully labeled public dataset for event-based non-intrusive load monitoring research," in *2nd Workshop on Data Mining Applications in Sustainability (SustKDD)*, 2011, p. 2012.

[19] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," 2011.

[20] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. V. Bajic, "Ampds: A public dataset for load disaggregation and eco-feedback research," in *Electrical Power & Energy Conference (EPEC), 2013 IEEE*. IEEE, 2013, pp. 1–6.

[21] S. Makonin, B. Ellert, I. V. Bajić, and F. Popowich, "Electricity, water, and natural gas consumption of a residential house in canada from 2012 to 2014," *Scientific data*, vol. 3, 2016.

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

[23] J. Z. Kolter and T. Jaakkola, "Approximate inference in additive factorial hmms with application to energy disaggregation," in *Artificial Intelligence and Statistics*, 2012, pp. 1472–1482.

[24] M. J. Johnson and A. S. Willsky, "Bayesian nonparametric hidden semi-markov models," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 673–701, 2013.

[25] M. Wytock and J. Z. Kolter, "Contextually supervised source separation with application to energy disaggregation." 2014.

[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[27] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[28] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[29] A. Marchiori, D. Hakkarinen, Q. Han, and L. Earle, "Circuit-level load monitoring for household energy management," *IEEE Pervasive Computing*, vol. 10, no. 1, pp. 40–48, 2011.

[30] L. K. Norford and S. B. Leeb, "Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms," *Energy and Buildings*, vol. 24, no. 1, pp. 51–64, 1996.

[31] M. L. Marceau and R. Zmeureanu, "Nonintrusive load disaggregation computer program to estimate the energy consumption of major end uses in residential buildings," *Energy conversion and management*, vol. 41, no. 13, pp. 1389–1403, 2000.

[32] K. D. Lee, S. B. Leeb, L. K. Norford, P. R. Armstrong, J. Holloway, and S. R. Shaw, "Estimation of variable-speed-drive power consumption from harmonic content," *IEEE Transactions on Energy Conversion*, vol. 20, no. 3, pp. 566–574, 2005.

[33] C. Laughman, K. Lee, R. Cox, S. Shaw, S. Leeb, L. Norford, and P. Armstrong, "Power signature analysis," *IEEE power and energy magazine*, vol. 99, no. 2, pp. 56–63, 2003.

[34] W. Wichakool, A.-T. Avestruz, R. W. Cox, and S. B. Leeb, "Modeling and estimating current harmonics of variable electronic loads," *IEEE Transactions on power electronics*, vol. 24, no. 12, pp. 2803–2811, 2009.

[35] H. Najmeddine, K. E. K. Drissi, C. Pasquier, C. Faure, K. Kerroum, A. Diop, T. Jouannet, and M. Michou, "State of art on load monitoring methods," in *Power and Energy Conference, 2008. PECon 2008. IEEE 2nd International*. IEEE, 2008, pp. 1256–1258.

[36] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *advances in neural information processing systems*, 2010, pp. 856–864.

[37] S. Mohamad, A. Bouchachia, and M. Sayed-Mouchaweh, "Asynchronous stochastic variational inference," *arXiv preprint arXiv:1801.04289*, 2018.

[38] S. Mohamad, M. Sayed-Mouchaweh, and A. Bouchachia, "Active learning for classifying data streams with unknown number of classes," *Neural Networks*, vol. 98, pp. 1–15, 2018.

[39] S. Mohamad, A. Bouchachia, and M. Sayed-Mouchaweh, "A bi-criteria active learning algorithm for dynamic data streams," *IEEE transactions on neural networks and learning systems*, 2016.