Programme Area: Smart Systems and Heat

Project: WP1 Appliance Disaggregation

Title: Pattern Mining

Abstract:

The purpose of this deliverable is to describes the second task which is related to pattern mining. Specifically, it presents an original approach for mining utility data usage patterns relying on a novel algorithm, Gaussian Latent Dirichlet Allocation (GLDA). A full empirical evaluation of the proposed algorithm using the ETI data is discussed highlighting its performance on various mining tasks.

Context:

The High Frequency Appliance Disaggregation Analysis (HFADA) project builds upon work undertaken in the Smart Systems and Heat (SSH) programme delivered by the Energy Systems Catapult for the ETI, to refine intelligence and gain detailed smart home energy data. The project analysed in depth data from five homes that trialed the SSH programme's Home Energy Management System (HEMS) to identify which appliances are present within a building and when they are in operation. The main goal of the HFADA project was to detect human behaviour patterns in order to forecast the home energy needs of people in the future. In particular the project delivered a detailed set of data mining algorithms to help identify patterns of building occupancy and energy use within domestic homes from water, gas and electricity data.

Bournemouth University

Project: HFADA

HIGH FREQUENCY APPLIANCE DISAGGREGATION ANALYSIS

Pattern Mining

Contents

1. History

Date	Issue	Details of Change
30/01/2018	Version 0.0	Initial Version. Authors: Saad Mohamad Chemseddine Mansouri Hamid Bouchachia

2. Documents Referenced

Ref	Document	Title
1	Word document that describes the HEMS data.	Data collection and data format – ELECTRIC, WATER and HEMS-V1 MONITORING
2	Word document that describes the HEMS V1 Mongo data base structure.	HEMS V1 Mongo Data Base Structure
3	Deliverable 1	HFADA_Deliverable_Ver2
4	Paper	Online Gaussian LDA for Unsupervised Pattern Mining from Utility Usage Data

3. Glossary of Terms

Ref	Description
ETI	Energy Technologies Institute
HEMS	Home Energy Management System (also referred to as HEMS V1)
HFAD	High Frequency Appliance Detection
LDA	Latent Dirichlet Allocation
GLDA	Gaussian Latent Dirichlet Allocation
NILM	Non-intrusive load monitoring

4. Executive Summary

This deliverable describes the second task which is related to pattern mining. Specifically, it presents an original approach for mining utility data usage patterns relying on a novel algorithm, Gaussian Latent Dirichlet Allocation (GLDA). A full empirical evaluation of the proposed algorithm using the ETI data is discussed highlighting its performance on various mining tasks.

5. Introduction

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 report, we propose a fully unsupervised NILM framework based on Bayesian hierarchical mixture models. In particular, we develop a new algorithm based on Gaussian Latent Dirichlet Allocation (GLDA) in order to extract global components (formally called topics) that summarise the energy signal. These components provide a representation of the consumption patterns. Designed to cope with big data, the GLDA algorithm, unlike existing NILM ones, does not focus on appliance recognition. To handle this massive data, it works online, so as new data arrive, the algorithm slef-adapt autonomously. Beside this novelty, compared to the existing NILM methods, the proposed one involves data emanating from different utilities (e.g, electricity, water and gas) and some sensor measurements. Finally, we propose different evaluation methods to analyse the results which show that GLDA finds useful patterns.

6. Pattern Mining

In the following we introduce mining approaching after providing some background related to human activity recognition and NILM techniques.

6.1 Context and Motivations

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 patterns of human behaviour. There are three main types of HAR approaches: sensor-based [1], vision-based [2] and radio-based [3]. A common feature of these methods is that they all require that the living environment be equipped with embedded devices (sensors) that emit data. On the other hand, non-intrusive

load monitoring (NILM) requires only a single meter per house or building that measures aggregated energy* 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 focuses not on extracting general human behaviour patterns but rather on identifying the appliances in use. This, however, can provide insight into the energy consumption behaviour of the residents and therefore can express user's 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 the proposed approach, 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 low-dimensional representation. This is possible due to the fact that the human behaves in a certain way following patterns. We are also lucky to have an extra-large amount of real world data which makes this approach more viable. Driven by this massive amount of data, our method operates online in real-time and is computationally efficient and scalable, unlike state-of-the-art probabilistic methods that posit detailed temporal relationships and involve complex inference steps.

6.2 NILM and Machine Learning

Since the earliest work on NILM [4], most NILM work has been based on signal processing and engineering approaches [5, 6] and 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 annotated data from individual appliances 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 installing 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 used for NILM [14] concentrate on disaggregating the whole house signal into its appliances' ones. In contrast, our unsupervised approach, as mentioned earlier, does not focus on identifying per-appliance signal. We instead propose a novel approach that extracts human behaviour patterns from home utility usage data. These patterns could be exploited for HAR as well as energy efficiency applications.

6.3 Analogy with Latent Dirichlet Allocation

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 (input data), global hidden variables, local hidden variables, and some fixed parameters are brought together to define the structure of the model. Under some

*Energy signal can be electricity consumption signal, water consumption signal, gas consumption signal or all of them together as in our case

assumptions, in addition to those indicated in [15], we end up with a Gaussian version of Latent Dirichlet Allocation (GDLA) where the observations (input data) are continuous. 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 therefore a mixture of components generating the input over a fixed period of time. In LDA, patterns are associated with documents that can be expressed by a 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. Nevertheless, this simplification leads to methods that are computationally more 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 to improve energy usage efficiency. In particular, understanding the usage and energy consumption patterns could be used to:

- predict the power demand (load forecasting),
- apply management policies and to avoid overloading the energy network and
- provide 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].

The mathematical details of the proposed Gaussian LDA are presented in [26][1].

[1] This paper was submitted to the journal of *Data Mining and Knowledge Discovery*.

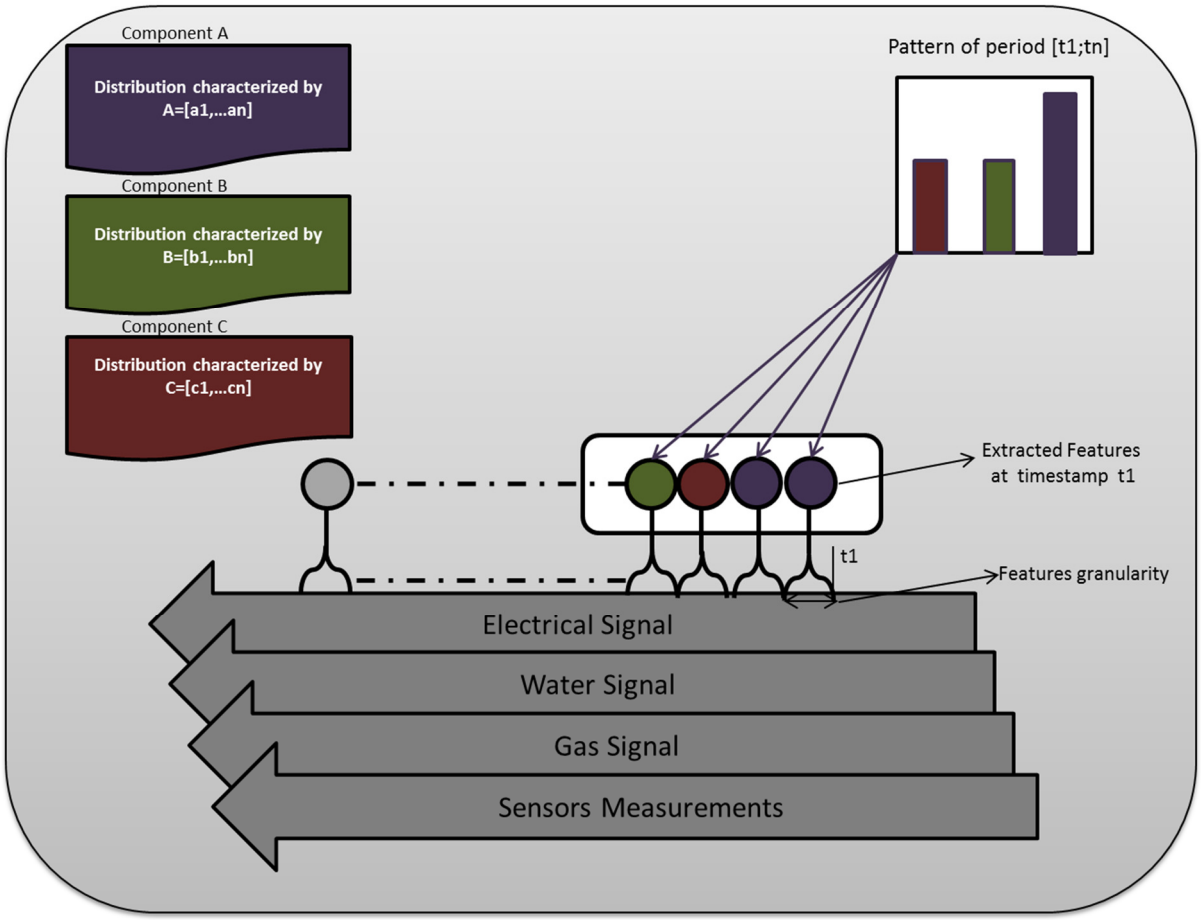


Figure 1: Elements of the proposed approach

6.4 Scalability

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 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 a big amount of data, an 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 technical details can be found in the enclosed paper [26]. Besides the advantage the data size offers, apart from [20, 21] whose sampling rate is very low, our data is the only one that includes water and gas usage data as well. Moreover, measurements provided by additional sensors are also exploited to refine the performance of the pattern mining algorithm. More details on the data can also be found in [26]. The diversity of the data is another motivation for adopting a pattern mining approach rather than traditional disaggregation approach.

7 Experiments

In this section we will present the dataset on which GLDA was tested, the evaluation methods used and the results obtained.

7.1 Data

The real-world multi-source utility usage data used here is provided by ETI. The data includes electricity signals (voltage and current signals) sampled at high sampling rate around 205 kHz, water and gas consumption sampled at low sampling rate. The data also contains other sensor measurements collected from the Home Energy Monitoring System (HEMS). In this study we will use 4Tb of utility usage data collected from house H71 over one month. Details on the data and how it is pre-processed can be found in Deliverable 1 [27].

7.2 Evaluation and Analysis

In order to evaluate GLDA, we use the perplexity measure. Perplexity is used to quantify the fit of the model to the data. 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 [22]. 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. This technique will also allow numerically checking the predicted consumption against the real consumption.

7.2.1 Model Fitness

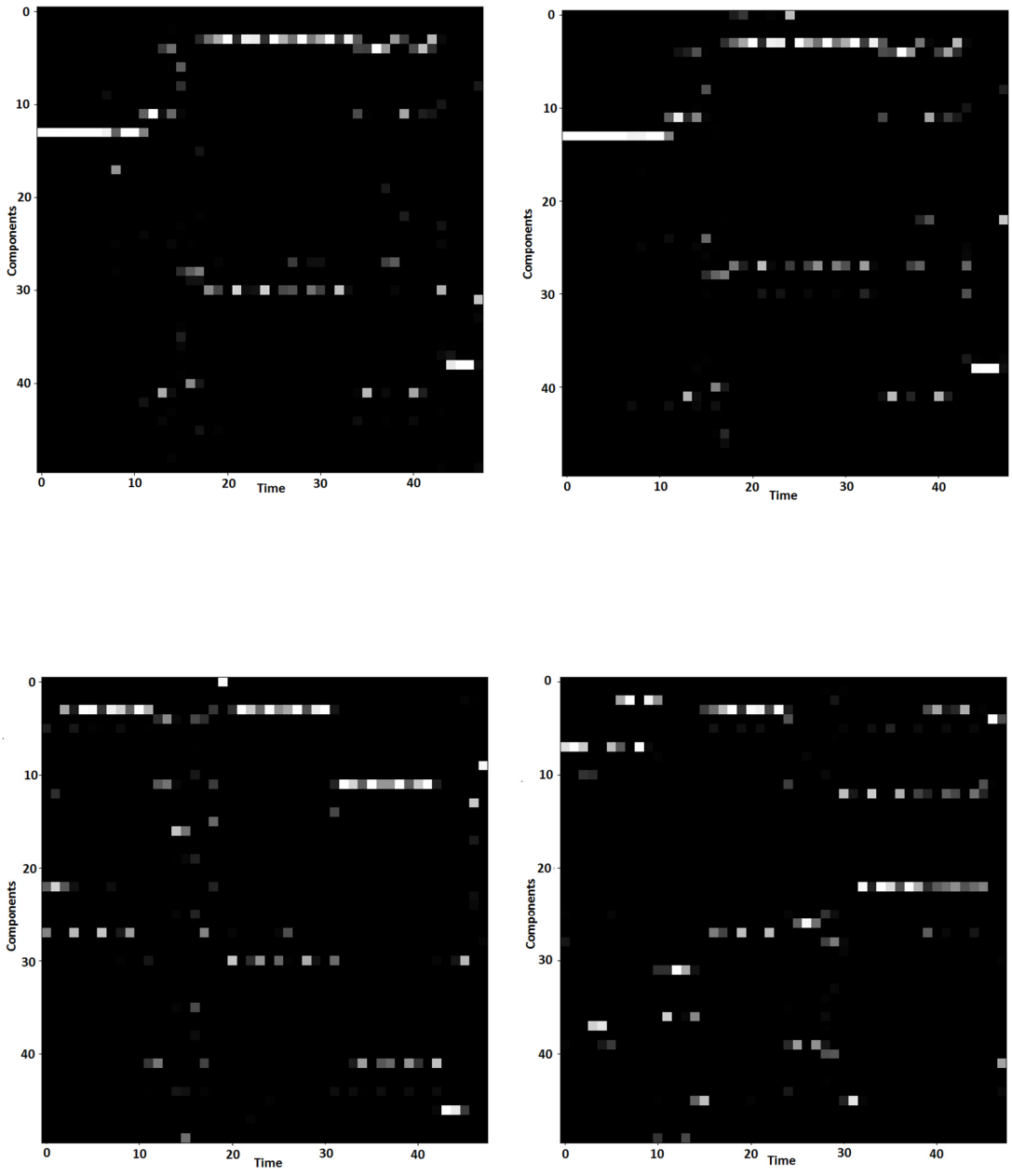
Although online GLDA converges for any valid hyper-parameters, the quality and speed of the convergence may depend on how these parameters are set. We run online GLDA on the training sets for different combination of parameters and take those that give the best the perplexity obtained on the validation set. Details can be found in the enclosed paper.

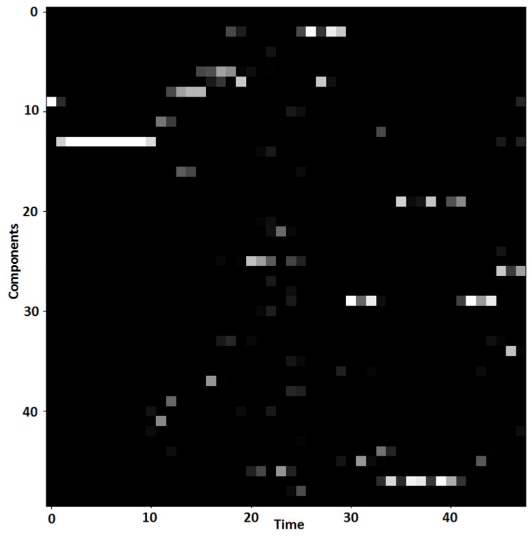
7.2.2 Pattern Regularity

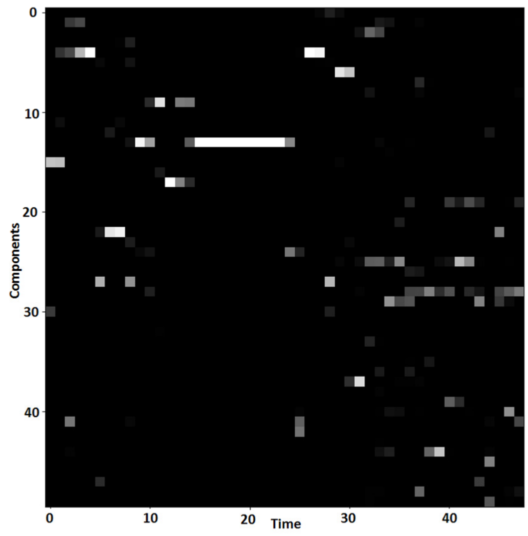
Using the optimal parameters' setting found in the previous experiments, we examine in the following the regularity of the mined patterns. To do that, we use the last two weeks of the data (from 18-05-2017 23:45:22 to 01-06-2017 23:45:22) for testing. To study the regularity of the energy consumption 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 columns. The left column corresponds to the week from 18-05-2017 23:45:22 to 26-05-2017 23:45:22. The right column 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 next to each other.

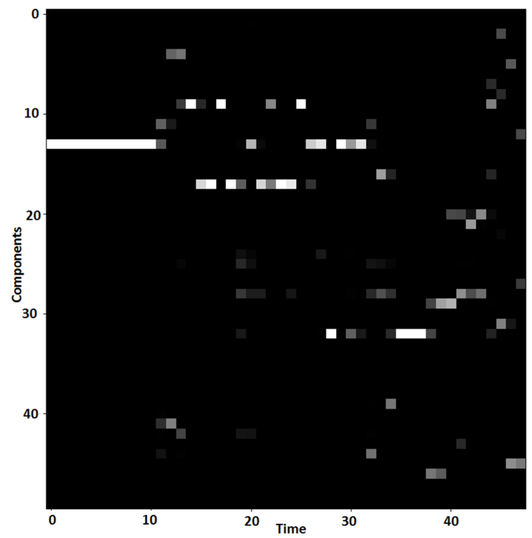
It can be clearly seen from these figures that there is regular patterns across columns. That is, similar energy consumption patterns appear across different weeks. Moreover, consumption patterns across working days within the same week are similar. On the other hand, for a specific week, the patterns over the weekend days and Friday are quite dissimilar

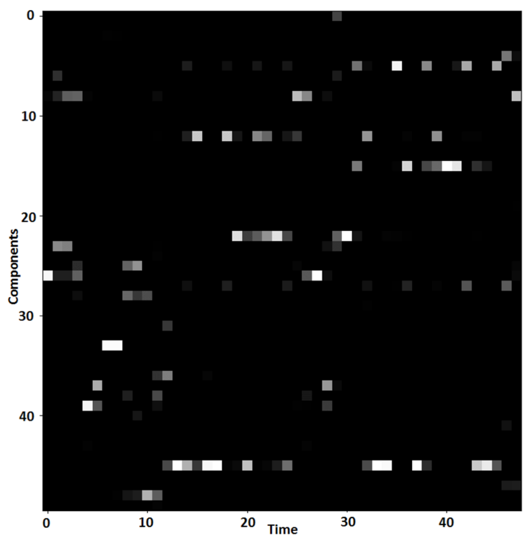
to the rest within and across that week. This regularity may be caused by regular user lifestyle leading to similar energy consumption behaviour within and across the weeks. Such regularity is violated in the weekend, where more random activities could take place. Note that the difference between the 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 energy consumption with each component, we can help validate the coherence of the extracted patterns and do forecasting.

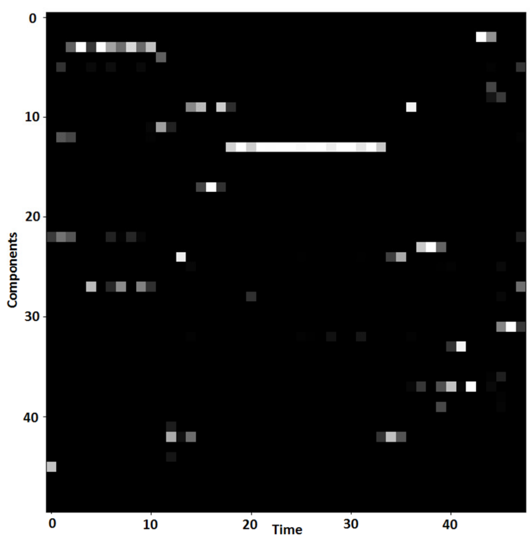


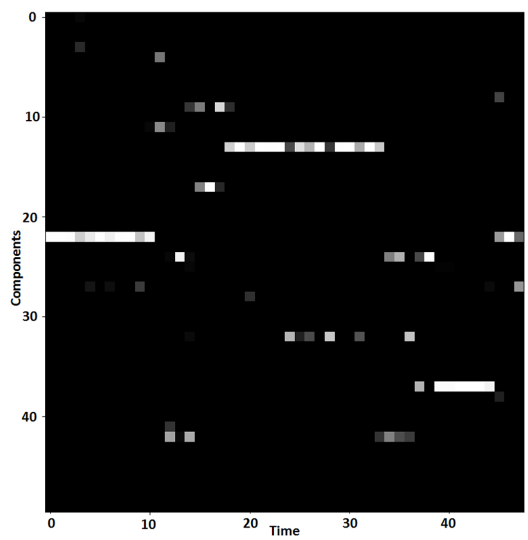

Components
0
10
20
30
40
Time
0
10
20
30
40


Components
0
10
20
30
40
Time
0
10
20
30
40


Components
0
10
20
30
40
Time
0
10
20
30
40


Components
0
10
20
30
40
Time
0
10
20
30
40


Components
0
10
20
30
40
Time
0
10
20
30
40


Components
0
10
20
30
40
Time
0
10
20
30
40

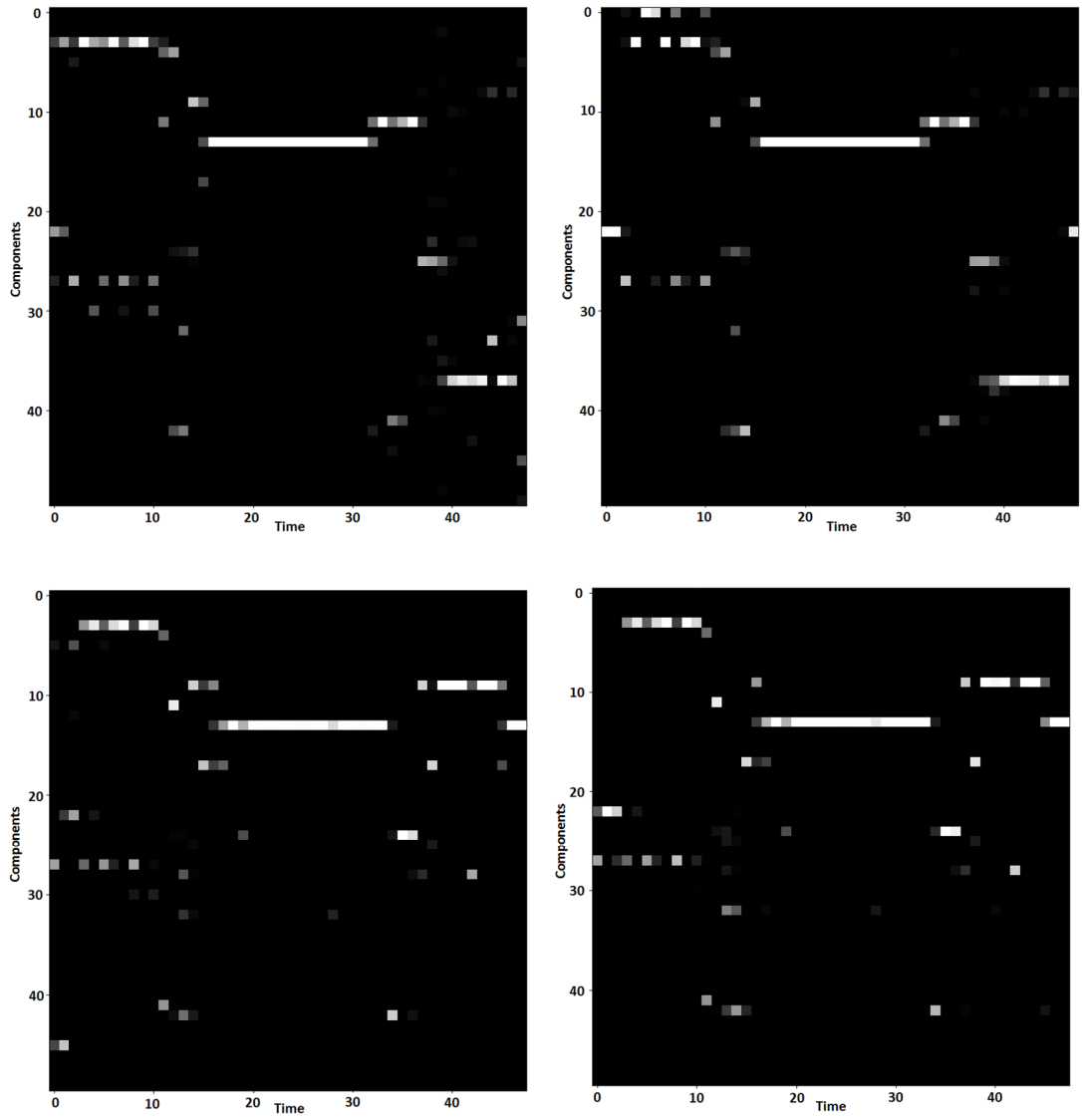


Figure 2: Emergence of patterns

7.2.3 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. We apply a

simple least-square regression method to map patterns to energy consumption. 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 and Fig. 4 are not fully the same.

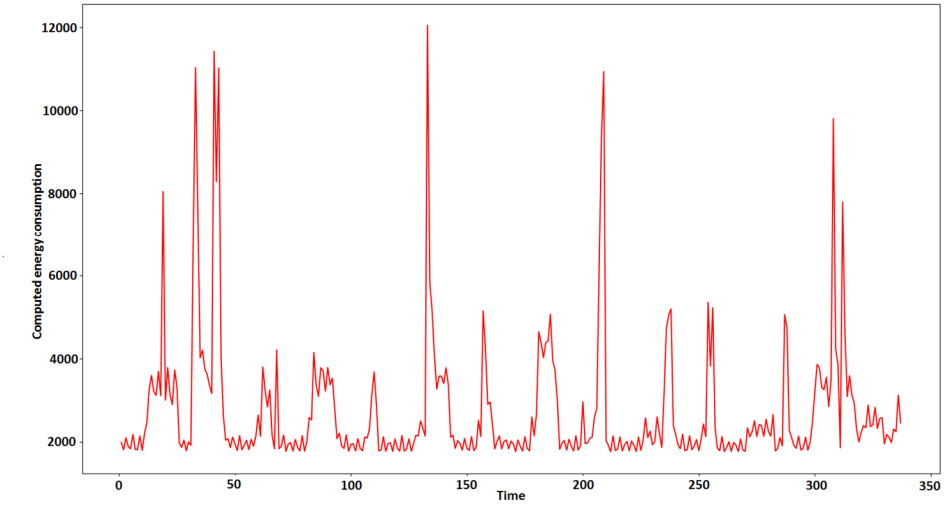


Figure 3: Computed energy consumption

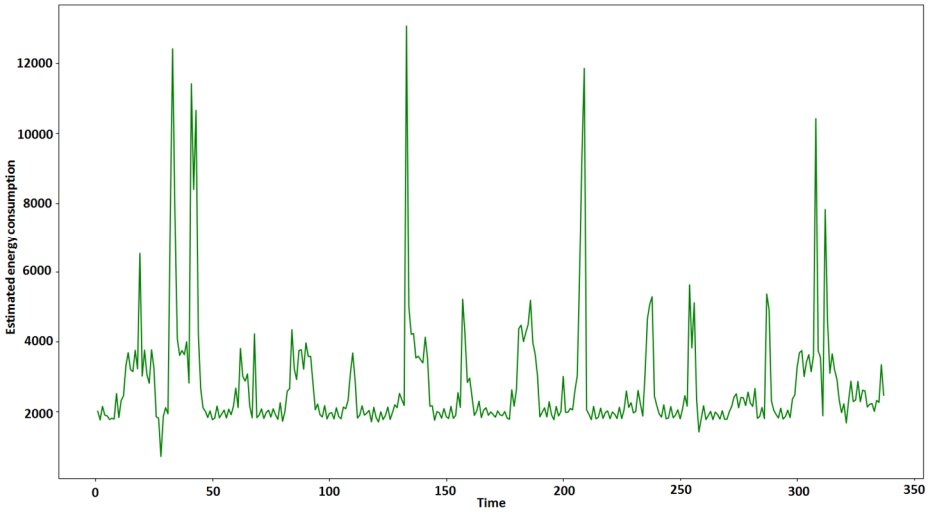


Figure 4: Estimated energy consumption

5. Conclusion

In this report, 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 sensor measurements. GLDA is fully unsupervised and works online which makes 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.

6. References

1. A. Bulling, U. Blanke, and B. Schiele, “A tutorial on human activity recognition using body-worn inertial sensors," ACM Computing Surveys (CSUR), 2014.

2. R. Poppe, “A survey on vision-based human action recognition," Image and vision computing, 2010.

3. S. Wang and G. Zhou, “A review on radio based activity recognition," Digital Communications and Networks, 2015.
4. G. W. Hart, “Nonintrusive appliance load monitoring," Proceedings of the IEEE, 1992.

5. M. Zeifman and K. Roth, “Nonintrusive appliance load monitoring: Review and outlook," IEEE transactions on Consumer Electronics, 2011.

6. A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, \Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey," Sensors, 2012.

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, 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,

9. D. Srinivasan, W. Ng, and A. Liew, “Neural-network-based signature recognition for harmonic source identi_cation," IEEE Transactions on Power Delivery, 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

11. A. G. Ruzzelli, C. Nicolas, A. Schoofs, and G. M. O'Hare, “Real-time recognition and pro_ling 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

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

13. Y.-X. Lai, C.-F. Lai, Y.-M. Huang, and H.-C. Chao, “Multi-appliance recognition system with hybrid svm/gmm classi_er in ubiquitous smart home," Information Sciences, 2013.

14. R. Bon_gli, 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

15. M. D. Ho_man, D. M. Blei, C. Wang, and J. Paisley, “Stochastic variational inference," The Journal of Machine Learning Research, 2013.

16. C. Fischer, “Feedback on household electricity consumption: a tool for saving energy?" Energy efficiency, 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, 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

21. S. Makonin, B. Ellert, I. V. Bajic, and F. Popowich, “Electricity, water, and natural gas consumption of a residential house in canada from 2012 to 2014," Scientific data, 2016.

22. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation," Journal of machine Learning research, 2003.

23. S. Mohamad, A. Bouchachia, and M. Sayed-Mouchaweh, “Asynchronous stochastic variational inference," arXiv preprint arXiv:1801.04289, 2018.

24. S. Mohamad, M. Sayed-Mouchaweh, and A. Bouchachia, \Active learning for classifying data streams with unknown number of classes," Neural Networks, 2018.

25. Mohamad, Saad, Abdelhamid Bouchachia, and Moamar Sayed-Mouchaweh. "A bi-criteria active learning algorithm for dynamic data streams." IEEE transactions on neural networks and learning systems (2016).

26. S. Mohamad, Chemseddine Mansouri, Abdelhamid Bouchachia. “Online Gaussian LDA for Unsupervised Pattern Mining from Utility Usage Data” submitted for ECML-PKDD 2018

27. S. Mohamad, Chemseddine Mansouri, Abdelhamid Bouchachia. “HFADA_Deliverable_Ver2” submitted to ETI on 20/11/2017