

Energy Disaggregation Using Principal Component Analysis Representation

Pierre V. DANTAS¹, Waldir SABINO S. Júnior² and Celso B. CARVALHO³

Federal University of Amazonas (UFAM), Amazonas, Brazil

Abstract. The main purpose of disaggregation is to decompose a signal into a set of other signals that together constitute it. This approach could be applied to audio signals, health care, home automation, ubiquitous systems and energy systems. It may be unworkable to individually measure the energy consumption of loads in a system simultaneously and, through disaggregation, we can make an inference using a main meter. The main contribution of this work is to use PCA to extract representativeness of an energy consumption signal we want to disaggregate, identifying its most relevant characteristics. The field of study is relevant because it allows information to be obtained in a simpler and cheaper way about the individual consumption of loads that make up a system. This opens up perspectives for other approaches such as smart grids and IoT. We demonstrate that when compared to other techniques, the proposal produces more accurate disaggregation results.

Keywords. energy disaggregation, principal component analysis (PCA), representativeness.

1. Introduction

Energy disaggregation consists of infer individual consumption of equipment that compose an amount of energy consumption, knowing only the latter. In this case, only one main meter is sufficient, which can make the measurement simpler and cheaper. In several situations, when we work with a database of great magnitude, may be more appropriate to work with the representativeness of a data set, which is a common condition in the area of signal processing [1].

In this work, we use *principal component analysis* (PCA) to extract representativeness from an energy consumption database and, thus, process disaggregation. We choose PCA in this work because it is widely used in multivariate analysis, used to transform a data set, with supposedly correlated variables, into a data set with uncorrelated variables and reduced dimensionality [2]. There are applications in data compression, extraction detection and pattern recognition [3, 4] and

¹Corresponding author: pierre.dantas@gmail.com

²waldirjr@ufam.edu.br

³ccarvalho_@ufam.edu.br

beyond [5, 6, 7]. PCA can be also analyzed as a projection issue where a vector in multidimensional space is projected in a vector subspace of smaller dimension, seeking to minimize the mean squared error or to maximize the variance of the projected vector. Although understanding that PCA has a higher computational cost operation, when compared to other techniques [8], we want to analyze the accuracy of disaggregation when dealing with representativeness. The increase in the accuracy of the disaggregation, in this case, may justify the use of a more complex previous processing.

We construct a new combination of initial variables in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information is compressed into the first components, that is, we obtain representativeness from energy consumption database to process disaggregation. So, the idea is PCA tries to put maximum possible information in the first component, then the maximum remaining information in the second and so on. Thus, the main idea is to assume that components of the highest energy of an energy signal corresponds to a *signature* (pattern). In this work, we observed that the six first principal components contain more than 99.9% of the signal energy, considering this the criterion for delimiting the dimension of the signal decomposition.

Regarding the database, we use the public Reference Energy Disaggregation Data Set (REED) database [9], prepared by Massachusetts Institute of Technology (MIT) for research purposes in energy disaggregation.

This article is organized as follows: first we present a related work section, detailing the advantages and disadvantages of techniques used in disaggregation. Then, we present a methodology section, detailing the procedures used to disaggregate an energy consumption signal using PCA, how the data were acquired and how we disaggregate the signal. Then, we detail the experiments, followed by a results section and, in the end, the discussions and conclusions of the work.

2. Related Work

There are several approaches to deal with energy disaggregation such as analyzing transitions in time domain, factorial hidden Markov models (FHMM), neural networks and machine learning algorithms. For a comprehensive survey and more details on non-intrusive load monitoring methods and techniques for energy disaggregation problem, reader should refer to a survey presented in [10]. We present next a summary of the main approaches.

2.1. Event and State-Based Approach

The approach to the problem of disaggregation had its first work published in 1992 [11], where the transitions of energy consumption signals in the time domain were analyzed. The event-based strategies center on the transition generated by appliances and use change detection algorithm to identify start and end of an event. Simple mean [12] calculates the mean squared error between the estimated energy consumption of each appliance and the consumption measured. Powerlets [12] uses a set of electrical energy consumption data from different appliances for training and identifies energy consumption patterns. Then, created a dictionary with the identified patterns and executed the disaggregation.

2.2. Hidden Markov Model

Hidden Markov model (HMM) is a technique whose each state is represented by a probability distribution function, modeling the observation corresponding to that state. Regardless of how HMM-based approaches were used in disaggregation [13, 14], specialized a priori knowledge is required. Consequently, its presentation is constrained by the way in which the models obtained seem approximate to true. In addition, HMM-based approaches have better success in controlling multi-state appliances, but their performance degrades for multi-state and variable uncontrolled appliances.

2.3. Graph Signal Processing

Graph signal processing (GSP) is an emerging field that extends the classical theory of signal processing to general graph indexed data. GSP is a powerful, scalable, and flexible approach to signal processing suited to machine learning and data mining issues. GSP-NILM disaggregates active power dataset without prior knowledge and relies on GSP to accomplish adaptive threshold, signal clustering and pattern matching [15, 16]. It works well if the average load of each appliance is sufficiently distinct from that of the other appliance load and if the power of each load does not fluctuate much. GSP approach requires that appliances be manually labeled after disaggregation.

2.4. Deep Learning

Deep learning is an approach to machine learning that has drawn on the human brain knowledge, statistics and applied mathematics. This is the artificial neural networks (ANN) that are made up of several layers. Recently, various deep learning architectures have been applied to the issue of energy disaggregation [17, 18, 19]. Some methods have been trained using few data, but in order to be well generalized, deep learning models need a lot of data. SSI-SD methodology [18] address disaggregation using data mining techniques and elaborate clusters (dictionaries) considering that the states of operation of the appliances and the signal of main consumption are dependent instances.

3. Methodology

The first stage of the methodology consists of elaborating a dictionary formed by the combining the energy consumption of the devices and the respective main consumption. This is a supervised learning where we use around 150,000 samples of energy consumption from each device of the energy consumption database [9]. In order to apply the PCA to extract representativeness, we performed the singular value decomposition (SVD) [2] of an array arranged from the energy consumption data via Hankel transformation [12]. With this transformation, we make it possible to treat a one-dimensional signal as a two-dimensional signal, in order to apply matrix operations. At the end of this operation, we obtain a matrix of eigenvectors arranged according to the energy of their eigenvalues.

Thus, we consider the eigenvalues with the highest signal energy to apply the concept of selection of principal components, that is, those that correspond to more than

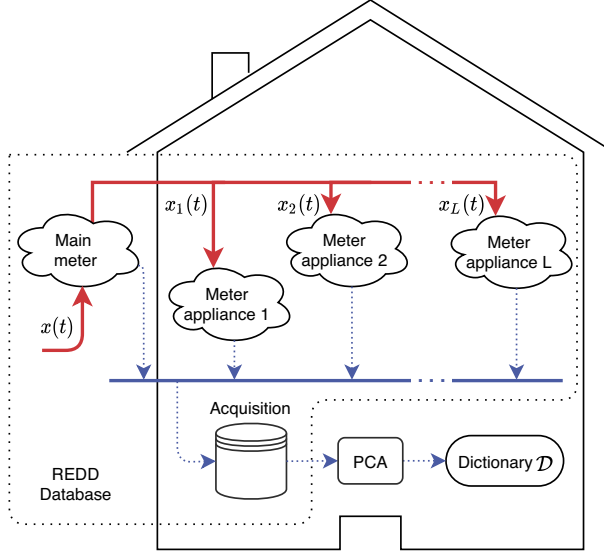


Figure 1. Framework for obtaining dictionary using PCA. The red line means electric power flow and the blue line means data flow. The blocks in the form of a cloud mean that we have a set of an appliance associated with an energy consumption meter.

99.9% of the total energy. Thus, we developed a dictionary composed of the eigenvalues and eigenvectors with the highest signal energy. We illustrate this situation in Figure 1.

After the training stage, we perform the second step of the signal disaggregation. At this point, we applied the PCA to a set of samples different from those selected for the training stage, in this case, considered as test samples, also around 150,000 samples. In this case, we only use the main electricity consumption data. We performed the PCA for each sample in the test set, obtaining a representation that allows comparison with the dictionary content obtained in the first step. Thus, we try to find, in the dictionary, the most similar set of this representation, called *matching* procedure [20], that consists on calculating dissimilarity, inferring about the solution of the disaggregation. With this, we estimate the most likely disaggregation set of energy appliances consumption. The estimated energy consumption signals of the L appliances are designated by $\hat{x}_{\{1,2,\dots,L\}}(t)$. We illustrate this situation in Figure 2.

4. Experiments

To evaluate the generalization of proposed method, we used *holdout validation*. We split up the data set into a *training* and *test* set. The training set is what the model is trained on, and the test set is used to see how well that model performs on unseen data. We divide the total data set in the ratio of 50% for training and 50% for testing. We decided to work with this proportion because we understand that, given the nature of the energy consumption signal, the number of samples is sufficient to allow a training stage with this proportion.

In the first stage of disaggregation, which consists in the elaboration of the dictionary, we use 150,000 samples of the main electrical energy consumption signal

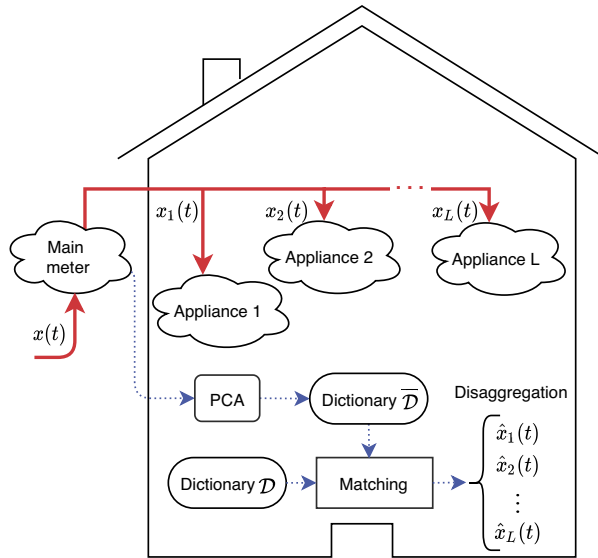


Figure 2. Framework for disaggregation of energy signal using PCA. The red line means electric power flow and the blue line means data flow. The blocks in the form of a cloud mean that we have an appliance.

and the same quantity for the electrical energy consumption signal of the appliances. Each signal was divided into 15,000 segments of dimension $w = 10$ to allow the application of Hankel's transformation. We apply this transformation of dimension $\ell = 5$. Then we apply singular value decomposition (SVD). We consider using the six principal components on PCA because we observed that this amount is sufficient to represent at least 99.9% of the signal energy. Then, for each segment of energy consumption signals, we select the first principal component and we elaborate the dictionary.

In the second stage of the disaggregation, we use 150,000 samples of a main electrical energy consumption signal only, divided into 15,000 segments of dimension $w = 10$, and we extracted the dictionary. Then, for each principal component, we look for a similarity in dictionary obtained in training step. Then, we recover the electricity consumption signals of the appliances, obtaining the estimates $\hat{x}_{\{1,2,\dots,L\}}(t)$.

This procedure was developed in MATLAB, after importing the energy consumption database. For the experiments, we used a computer with a 1.7GHz Intel Core i5 processor, 2 cores, 4GB memory, 1,600MHz DDR3, macOS Sierra operating system version 10.12.5.

5. Results

In Table 1, we present the accuracy of disaggregation obtained with the application of PCA. Note that the accuracy obtained is greater than the other methodologies used in comparison. We were careful to compare the results with other methodologies that also used the REDD database to implement the disaggregation. Likewise, the metric for calculating the disaggregation is the same in all works. Note that there are several

parameters that influence the accuracy of the disaggregation and we found that, given the same methodology, the result of the accuracy may be different.

Note that the accuracy is calculated at the end of the disaggregation of all samples in the test data set. In order to determine accuracy, it is necessary to know the true solution of the disaggregation. In this case, the database used allows this possibility, given that this database was specifically designed for these purposes. In a different case, it would not be possible to calculate the disaggregation accuracy and we would have to use the performance parameter of the test step to rate the effectiveness of the training. In these cases, it may be necessary to use unsupervised learning methods [21, 22] given that it does not require any labeled data.

Table 1. Accuracy of methodologies of energy disaggregation.

Methodology	Accuracy (%)
PCA	97,2
SSI-SD	87,3
Powerlets	79,0
FHMM	59,6
Simple Mean	39,0

6. Conclusion

A new approach for disaggregating energy consumption signals by a non-intrusive method is proposed. The estimate of the individual energy consumption of devices in a household, given the knowledge of general consumption only, is based on the extraction of representativeness, the elaboration of a dictionary and the search for a solution of disaggregation by metrics of dissimilarity analysis. The approach is based on the application of principal component analysis (PCA) to develop dictionaries that represent the operating states of the devices that make up the general energy consumption. This allows the elaboration of the dictionary based on the analysis of the magnitude of the signal energy.

The proposed algorithm was tested using a data set specifically made for the purpose of studies on the disaggregation. The result showed better performance, on average, for other methodologies that used the same database and the same accuracy calculation metrics. Emphasizing that this estimation method is supervised. The advantages of the proposed algorithm include obtaining greater disaggregation accuracy at the expense of a higher computational cost, given the required matrix decomposition operations. This method is, therefore, a new highly accurate disaggregation method.

7. Acknowledgment

This research, as provided for in Article 48 of Decree No. 6008/2006, was funded by Samsung Electronics of Amazonia LTDA, under the terms of Federal Law No. 8387/1991, through agreement No. 004, signed with the Center for R&D in Electronics and Information from the Federal University of Amazonas - CETELI/UFAM.

References

- [1] Elhamifar E, Sapiro G, Sastry SS. Dissimilarity-based sparse subset selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2016;38(11):2182–2197.
- [2] Syme C. Principal components analysis. In: *Encyclopedia of Ecology*. Elsevier; 2018. p. 566–573.
- [3] Harris T, Yuan H. Filtering and frequency interpretations of singular spectrum analysis. *Physica D: Nonlinear Phenomena*. 2010;239(20–22):1958–1967.
- [4] Roy Chowdhury M, Tripathi S, De S. Adaptive Multivariate Data Compression in Smart Metering Internet of Things. *IEEE Transactions on Industrial Informatics*. 2020;PP:1–1.
- [5] Qi J, Jiang G, Li G, Sun Y, Tao B. Surface EMG hand gesture recognition system based on PCA and GRNN. *Neural Computing and Applications*. 2019;p. 6343–6351.
- [6] Jafarzadegan M, Safi F, Beheshti Z. Combining Hierarchical Clustering approaches using the PCA Method. *Expert Systems with Applications*. 2019;137.
- [7] Ait-Sahalia Y, Xiu D. Principal Component Analysis of High-Frequency Data. *Journal of the American Statistical Association*. 2017;114(525):287–303.
- [8] Tkachenko R, Izonin I. Model and Principles for the Implementation of Neural-Like Structures Based on Geometric Data Transformations. In: *Advances in Computer Science for Engineering and Education*. Springer; 2019. p. 578–587.
- [9] Kolter J, Johnson M. REDD: A Public Data Set for Energy Disaggregation Research. *Artif Intell*. 2011;25.
- [10] Faustine A, Mvungi N, Kaijage S, Kisangiri M. A Survey on Non-Intrusive Load Monitoring Methodies and Techniques for Energy Disaggregation Problem. *CoRR*. 2017;abs/1703.00785.
- [11] Hart G. Nonintrusive Appliance Load Monitoring. *Proceedings of the IEEE*. 1992 01;80(12):1870 – 1891.
- [12] Elhamifar E, Sastry S. Energy Disaggregation via Learning ‘Powerlets’ and Sparse Coding. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press; 2015. p. 629–635.
- [13] Kolter JZ, Jaakkola T. Approximate inference in additive factorial HMMs with application to energy disaggregation. In: *Journal of Machine Learning Research*; 2012. .
- [14] Yang C, Wu Z. Research on Non-intrusive Load Decomposition Based on FHMM. *IOP Conference Series: Materials Science and Engineering*. 2020;768:62046.
- [15] Stankovic V, Liao J, Stankovic L. A graph-based signal processing approach for low-rate energy disaggregation. In: *2014 IEEE symposium on computational intelligence for engineering solutions (CIES)*. IEEE; 2014. p. 81–87.
- [16] Li D, Dick S. A graph-based semi-supervised learning approach towards household energy disaggregation. In: *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE; 2017. p. 1–7.
- [17] Shin C, Rho S, Lee H, Rhee W. Data requirements for applying machine learning to energy disaggregation. *Energies*. 2019;12:201–213.
- [18] Dantas P, Junior W. Energy Disaggregation via Data Mining. In: *Brazilian Technology Symposium*. Springer; 2019. p. 541–546.
- [19] Kelly J, Knottenbelt W. Neural NILM: Deep neural networks applied to energy disaggregation. In: *BuildSys 2015 - Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built*; 2015. .
- [20] Hakak S, Kamsin A, Palaiahnakote S, Gilkar G, Khan W, Imran M. Exact String Matching Algorithms: Survey, Issues, and Future Research Directions. *IEEE Access*. 2019;PP:1–1.
- [21] Holweger J, Dorokhova M, Bloch L, Ballif C, Wyrsh N. Unsupervised algorithm for disaggregating low-sampling-rate electricity consumption of households. *Sustainable Energy, Grids and Networks*. 2019;19:100244.
- [22] Hosseini S, Kelouwani S, Agbossou K, Cardenas A, Henao NF. Adaptive on-line unsupervised appliance modeling for autonomous household database construction. *International Journal of Electrical Power & Energy Systems*. 2019;112:156–168.