Nuclear Norm Regularized Robust Dictionary Learning for Energy Disaggregation

Megha Gupta
IIIT-Delhi
meghag@iiitd.ac.in

Angshul Majumdar
IIIT-Delhi
angshul@iiitd.ac.in

***Abstract*— The goal of this work is energy disaggregation. A recent work showed that instead of employing the usual Euclidean norm cost function for dictionary learning, better results can be achieved by learning the dictionaries in a robust fashion by employing an l_1-norm cost function; this is because energy data is corrupted by large but sparse outliers. In this work we propose to improve the robust dictionary learning approach by imposing low-rank penalty on the learned coefficients. The ensuing formulation is solved using a combination of Split Bregman and Majorization Minimization approach. Experiments on the REDD dataset reveal that our proposed method yields better results than both the robust dictionary learning technique and the recently published work on powerlet energy disaggregation.**

Keywords*— *Energy Disaggregation, Dictionary Learning, Robust Learning

I. INTRODUCTION

Currently, residential and commercial buildings account for 40% of the total energy consumption [1]. Studies have estimated that 20% of this consumption could be avoided with changes in user behavior [2]. Energy disaggregation is the task of segregating the combined energy signal of a building into the energy consumption of individual appliances. Disaggregation presents a way in which information regarding consumption patterns of individuals can be fed back to consumers with the goal of increasing their awareness about energy usage and its wastage. Studies have shown that such precise and detailed feedback to consumers can be quite effective towards improving energy conservation [3].

The approach towards energy disaggregation is broadly based on the nature of the targeted household and commercial appliances. These appliances can be broadly categorized as simple two-state (on/off) appliances such as electrical toasters and irons; more complex multistate appliances like refrigerators and washing machines; and continuously varying appliances such as IT loads (printers, modems, laptops etc.). The earliest techniques were based on using real and reactive power measured by residential smart meters. The appliances' power consumption patterns were modelled as finite state machines [4]. These techniques were successful for desegregating simple two state and multistate appliances, but they performed poorly in the case of time-varying appliances, which do not show a marked step increase in the power. More recent techniques, based on stochastic finite state machines (Hidden Markov Models) [5], have improved upon the prior approach. Current techniques are based on learning a basis / model for individual appliances. Sparse coding and dictionary learning based approaches like [6, 7] fall under this category. Given the limitations in space it is not possible to discuss all the prior studies in this area in detail; the interested reader should peruse [8].

Energy disaggregation has two phases. In the training phase, the power consumption data for each appliance is collected separately over time, and the model for the appliance is built. In the disaggregation phase, the composite data for multiple appliance is available, and the task is to estimate the power consumed by each appliance.

In this work we propose to improve the dictionary learning based techniques. Usually the dictionaries are learnt by minimizing the l_2-norm; this is mainly because it has a closed form solution (easy to minimize). It is well known that the Euclidean norm is sensitive to outliers.

Dictionary learning based methods learn a codebook that can represent the training data (X). This is expressed as:

$$X_i = D_i Z_i + \varepsilon,\ i = 1...\mathrm{N} \tag{1}$$

Here N is the number of appliances, D is the codebook / dictionary and the Z is the coefficient. Both D and Z need to be learnt.

All prior studies assumed that the modelling error (ε) is small and Normally distributed. Therefore, a Euclidean norm based minimization technique was employed. In a generic fashion, the learning can be expressed as:

$$\min_{D_i, Z_i} \left\| X_i - D_i Z_i \right\|_F^2 + R(D_i, Z_i) \tag{2}$$

where R is some penalty on D and Z. The main cost function is Euclidean hinged on the assumption $\varepsilon \sim N(0, \sigma^2)$.

However, the assumption that the modelling error is Normally distributed is incorrect. Training data always show spikes that are not typical to the appliance under study. They arise out of power surges or from transients. In such a case the modelling error does not follow a Normal distribution. It is large in magnitude but sparse. In this work, we address the more realistic noise model.

During disaggregation, all prior studies assume that the total power logged by the smart-meter is a sum of the power consumed by individual appliances that are turned on. This is expressed as:

$$X = \sum_i X_i + \varepsilon \tag{3}$$

Here too, the assumption is that the modelling error is small and approximately follows a Normal distribution. There are two reasons why this assumption is wrong. The first one has already been discussed. There are unforeseen spikes owing to uncontrollable reasons.

The other reason is owing to non-linear effects. The assumption that total power is a sum of individual power consumed by different appliances only holds for passive loads. The non-linearity can arise in two ways –

1. Today, most of our appliances such as refrigerators, AC's, washers, microwaves, laptops, printers etc. are quite sophisticated and cannot be modeled as passive loads. They have internal sources of electromagnetic emission, e.g. the switched mode power supplies (SMPS) in a desktop or a laptop adapter. These secondary sources of emission can interfere with the loads on the power lines depending on the proximity of the loads as well as their frequency response. In such cases, where the appliance needs to be modeled as a combination of a source and a load, the linear mixing model does not hold [9].

2. Secondly, the reactive components of the loads (transformers, magnetic and capacitive elements within the power supplies and AC to DC converters) exhibit non-linear behavior depending on the frequency of operation.

Non-linear loads are however, challenging to model. A seemingly unrelated of research, in hyperspectral unmixing, faces a similar issue. A recent study [10], showed that the non-linear mixing problem can be approximated as a sum of linear mixing and non-linear perturbations. The perturbations can be assumed to be sparse, i.e. the effect is localized but may be of relatively large magnitude.

In a recent work [11], a similar approach is followed – it assumed that the linear mixing model for energy holds in most cases (since this model is known to yield good results for simple loads), but that there are a few large perturbations arising out of the inherent non-linearity inside the load. The reasoning here is that linearity holds at the power line frequency (50\60Hz) while the non-linearity arises from the electromagnetic emission within the appliances at higher frequencies. In order to meet federal regulations on emissions, some of these appliances are fitted with good quality filters that restrict the emission. Secondly, these emissions are likely to decay with the length of the transmission cables. Therefore, a perturbation based model of the load non-linearities may be suitable in this context.

In short, we argue that both during training and during actual disaggregation, the assumption that the modelling error is small (Normally distributed) is not true. In both cases the actual noise is sparse but of large magnitude. During training, the sparse noise arises out of power surges; during disaggregation it arises from power surges and other non-linear effects. Since prior studies assumed the noise to be small, they employed an l_2-norm cost function. It is well known that the Euclidean norm is not robust to outliers (large and sparse perturbations). Therefore in this work, we propose to employ an l_1-norm cost function for both training and disaggregation. This makes the cost function robust to outliers.

In (2), we did not mention any particular penalty. Usually the penalties are supervised so as to maximize discrimination [6, 7]. In this work we propose a new penalty term on the coefficients (Z), based on rank-deficiency. The details will be discussed later. Results show that a combination of rank deficiency and robust learning produces results that are better than more sophisticated state-of-the-art techniques.

The paper is organized in several sections. Relevant studies are discussed in section II. The proposed method is described in section III. The experimental results are shown in section IV. The conclusions of this work are discussed in section V.

II. Literature Review

Kolter et al [6], assumed that there is training data collected over time, where the smart-meter logs only consumption from a single device only. This can be expressed as X_i where i is the index for an appliance, the columns of X_i are the readings over a period of time. For each appliance they learnt a codebook as shown in (1), where D_i represents the codebook/dictionary and Z_i are the loading coefficients, assumed to be sparse. This is a typical dictionary learning problem with sparse coefficients. In [6] various penalties were proposed on the dictionaries and the coefficients. In the simplest formulation the problem is solved via:

$$\min_{D_i, Z_i} \left\| X_i - D_i Z_i \right\|_F^2 + \lambda \left\| Z_i \right\|_1 \tag{4}$$

This is bi-linear problem; it is usually solved via alternating minimization, i.e. the coefficients are estimated assuming the dictionary is fixed; and the dictionary / codebook is updated assuming that the coefficients are fixed. Off-the-shelf algorithms exist for each of the two problems. Usually, the atoms of the dictionary are normalized to prevent degenerate solutions.

Learning the dictionary constitutes the training phase. During actual operation, several appliances are likely to be in use simultaneously. They [6] make the assumption that the aggregate reading by the smart-meter is a sum of the powers for individual appliances. Thus if X is the total power from N appliances (where the columns indicate smart-meter readings over the same period of time as in training) the aggregate power is modeled by (3). By imputing (1) in (3), one can express (3) as –

$$X = [D_1 | \ldots | D_N] \begin{bmatrix} Z_1 \\ \ldots \\ Z_N \end{bmatrix} \tag{5}$$

The loading coefficients can be solved using l_1-norm minimization. Once the loading coefficients are estimated, the consumption for each appliance is obtained by:

$$\hat{X}_i = D_i Z_i,\ i = 1...\mathrm{N} \tag{6}$$

Prior studies in dictionary learning are based on minimizing an l_2-norm data mismatch – the underlying assumption being that the noise (ε) is approximately Normally distributed. This does not hold for our problem; the reasons have been explained in the introduction. The noise is large but sparse. A more appropriate model would be one where ε is sparse (modelling non-linear perturbations). If there are such large outliers the estimate from Euclidean norm minimization is skewed towards the outlier. For a robust estimate [11] proposed to replace the l_2-norm by an l_1-norm data mismatch (since ε is sparse).

$$\min_{D_i, Z_i} \|X_i - D_i Z_i\|_1 + \lambda \|Z_i\|_1 \tag{7}$$

Both [6] and [11] imposed sparsity constraint on the coefficients. However we fail to see any clear motivation behind the requirement of enforcing sparsity on the loading coefficients.

Once the dictionary is learnt, [11] follows the procedure similar to [6]. The aggregate consumption (X) is assumed to the sum of consumptions from individual appliances (X_i's). However [11] acknowledges that the modelling error ε is large but sparse. Thus they estimated loading coefficients (Z_i's) by solving,

$$\min_{Z_i} \left\| X - \sum_i D_i Z_i \right\|_1 + \lambda \sum_i \|Z_i\|_1 \tag{8}$$

To the best of our knowledge there is hardly any study on robust dictionary learning (apart from [11]). There is a large body of literature in robust statistics that argues against the usage of l_2-norm minimization; it works when the deviations are small – approximately Normally distributed; but fail when there are large outliers (as in our case). The Huber function [12] has been in use for more than half a century in this respect. The Huber function is an approximation of the more recent absolute distance based measures (l_1-norm). Recent studies in robust estimation prefer minimizing the l_1-norm instead of the Huber function [12]-[14]. The l_1-norm does not bloat the distance between the estimate and the outliers and hence is robust.

The problem with minimizing the l_1-norm is computational. However, over the years various techniques have been developed. The earliest known method is based on Simplex [15]; Iterative Reweighted Least Squares [16] used to be another simple yet approximate technique. Other approaches include descent based method introduced by [17] and Maximum Likelihood approach [18].

In [11] a more modern approach is followed for solving the l_1-norm cost function. It is based on variable splitting and augmented Lagrangian. It decomposes the difficult problem of l_1-norm data mismatch into several easier sub-problems whose solutions exist.

III. Proposed Approach

In prior studies [6, 11] the loading coefficients were assumed to be sparse. However, sparsity does not model any reasonable aspect of the disaggregation problem; it only learns a dictionary to express the smart-meter signals in a sparse fashion.

Let us take a closer look at the problem; especially the training phase. The x_i's basically tell us the power consumption of individual devices across time. Ideally they are non-zero only when they are ON and zero when OFF. Consider a utopian situation where the devices are turned on exactly at the same time every day; in that case the matrix formed by stacking the x_i's will be of rank-1; since all the columns x_i's will be the same. In general this assumption will never hold in practice, each of the x_i's will be time shifted versions of each other. Here we propose to learn a dictionary that approximately aligns the input signals (x_i's) so that the resultant output (coefficients z_i's) are approximately aligned. If the z_i's are approximately aligned the coefficients matrix Zi (formed by stacking z_i's as columns) will be of low-rank. Following this assumption, we learn the dictionaries such that the resultant coefficients will be of low-rank. This is formally expressed as,

$$\min_{D_i, Z_i} \|X_i - D_i Z_i\|_1 + \lambda \|Z_i\|_{NN} \tag{9}$$

The nuclear norm is the convex surrogate of the rank of a matrix; it enforces rank deficiency on the variable.

For disaggregation, we follow the standard superposition model –

$$X_i = \sum_i X_i + \varepsilon = \sum_i D_i Z_i + \varepsilon$$

The dictionaries are learnt in the training phase; during disaggregation, we propose solving the following problem.

$$\min_{Z_i} \left\| X - [D_1 | ... | D_N] \begin{bmatrix} Z_1 \\ ... \\ Z_N \end{bmatrix} \right\|_1 + \lambda \sum_i \|Z_i\|_{NN} \tag{10}$$

A. Deriving a solution for Training Phase

Solving the robust dictionary learning problem, subject to nuclear norm penalty is new. Here we adopt the Split Bregman approach. We substitute $P = X - DZ$ (dropping the subscripts for notational simplicity) and introduce a Bregman relaxation variable B. This leads to the following formulation of (9),

$$\min_{D,Z,P} \|P\|_1 + \lambda \|Z\|_{NN} + \mu \|P - X + DZ - B\|_F^2 \tag{12}$$

Alternating minimization of (12) leads to the following sub-problems:

$$\mathrm{P1}: \min_D \|P - X + DZ - B\|_F^2 \tag{13a}$$

$$\mathrm{P2}: \min_Z \lambda \|Z\|_{NN} + \mu \|P - X + DZ - B\|_F^2 \tag{13b}$$

$$\text{P3}: \min_{P} \|P\|_1 + \mu \|P - X + DZ - B\|_F^2 \quad (13c)$$

Solving P1 is straightforward, it is a least squares problem with analytic solution. P2 is a nuclear norm regularized least squares minimization problem. This is efficiently solved using singular value shrinkage [19, 20]. The last sub-problem P3, also has a closed form solution in the form of soft thresholding [21].

B. Deriving a solution for Disaggregation Phase

We follow the Split Bregman approach here as well. We make the substitution $P = X - DZ$ where $D = [D_1 | ... | D_N]$ and $Z = \begin{bmatrix} Z_1 \\ ... \\ Z_N \end{bmatrix}$. After introducing the Bregman relaxation variable, the problem (10) takes the form,

$$\min_{Z,P} \|P\|_1 + \mu \|P - X + DZ - B\|_F^2 + \lambda \sum_i \|Z_i\|_{NN} \quad (14)$$

As before this can be decomposed into the following sub-problems:

$$\text{P1}: \min_{P} \|P\|_1 + \mu \|P - X + DZ - B\|_F^2 \quad (15a)$$

$$\text{P2}: \min_{Z_i} \mu \|P - X + DZ - B\|_F^2 + \lambda \sum_i \|Z_i\|_{NN} \quad (15b)$$

We have already discussed the solution for P1; it is solved via soft thresholding. Solving P2 is however tricky. To solve this, we need to decouple the problem. This can be done via Majorization-Minimization (MM) [21]. We represent P2 in a simpler fashion.

$$\|Y - DZ\|_F^2 + \eta \sum_i \|Z_i\|_{NN} \quad (16)$$

where $Y = X + B - P$ and $\eta = \lambda/\mu$.

In every iteration, MM of (16) leads to decoupling –

$$\|T - Z\|_F^2 + \lambda \sum_i \|Z_i\|_{NN} \quad (17)$$

where $T = Z_{k-1} + \frac{1}{\alpha} D^T \left(Y - DZ_{k-1}\right)$; α is the maximum eigenvalue of $D^T D$.

The decoupled problem is easy to solve; it can be segregated for every appliance 'i' in the following way,

$$\min_{Z_i} \|T_i - Z_i\|_F^2 + \lambda \|Z_i\|_{NN} \quad (18)$$

This is solved by one step of singular value shrinkage [19, 20].

C. Updating the Bregman Relaxation Variable

The final step is to update the relaxation variable B for all the problems. This is done by simple gradient descent.

$$B \leftarrow P - X + DZ - B \quad (19)$$

There are two stopping criteria for the Split Bregman algorithm. Iterations continue till the objective function converges (to a local minima). The other stopping criterion is a limit on the maximum number of iterations. We have kept it to be 200.

IV. Experimental Results

Our algorithm requires specifying the parameter λ and the hyperparameter μ. Some recent studies have shown that in a Split Bregman based technique, one can put $\lambda=1$ and only tune the μ. We use the simple L-curve method [22] for tuning the hyper-parameter. The number of dictionary atoms used is 144. The dictionary atoms are initialized by randomly picking up columns from the training data.

For energy disaggregation, we report results on the popular REDD [23] dataset. The dataset consists of power consumption signals from six different houses, where for each house, the whole electricity consumption as well as electricity consumptions of about twenty different devices are recorded. The signals from each house are collected over a period of two weeks with a high frequency sampling rate of 15kHz. In the standard evaluation protocol, the 5th house is omitted since it does not have enough data.

Table. 1. Description of Appliances in Houses

House	Appliances
1	Electronics, Lighting, Refrigerator, Disposal, Dishwasher, Furnace, Washer Dryer, Smoke Alarms, Bathroom GFI, Kitchen Outlets, Microwave
2	Lighting, Refrigerator, Dishwasher, Washer Dryer, Bathroom GFI, Kitchen Outlets, Oven, Microwave, Electric Heat, Stove
3	Electronics, Lighting, Refrigerator, Disposal, Dishwasher, Furnace, Washer Dryer, Bathroom GFI, Kitchen Outlets, Microwave, Electric Heat, Outdoor Outlets
4	Lighting, Dishwasher, Furnace, Washer Dryer, Smoke Alarms, Bathroom GFI, Kitchen Outlets, Stove, Disposal, Air Conditioning
6	Lighting, Refrigerator, Disposal, Dishwasher, Washer Dryer, Kitchen Outlets, Microwave, Stove

The disaggregation accuracy is defined as follows [23] –

$$Acc = 1 - \frac{\sum_t \sum_n \left| \hat{y}_t^{(i)} - y_t^{(i)} \right|}{2 \sum_t \bar{y}_t} \quad (20)$$

where t denotes time instant and n denotes a device; the 2 factor in the denominator is to discount the fact that the absolute value will "double count" errors.

In the previous work [11] it was already shown that simple robust dictionary learning yields better results than the well known Factorial Hidden Markov Model (FHMM)

[23] and sophisticated methods like discriminative sparse coding [6]. In this work, we show that our results are better than Robust Dictionary Learning (Robust DL) [11] (this will empirically prove that ours is better than [6] and [23]). We also compare our results with the state-of-the-art PED [7] (published in 2015). The comparative results are shown in Table 2.

As outlined in [23], there are two modes of testing. The first mode is simple, a portion of the data from every household is used as training samples and rest (from those households) is used for prediction. The second mode is more challenging, the data from four households are used for training and the remaining one is used for prediction; this is a more challenging problem.

Table. 2. Energy Disaggregation Results (in %)

House	Mode 1			Mode 2		
	Robust DL	PED	Prop	Robust DL	PED	Prop
1	70.1	81.6	**82.8**	52.1	46.0	**54.2**
2	61.9	79.0	**79.7**	55.7	49.2	**58.4**
3	61.0	61.8	**65.9**	43.3	31.7	**48.9**
4	71.0	58.5	**73.5**	59.8	50.9	**63.8**
6	64.7	79.1	**79.9**	60.0	54.5	**64.0**

Robust dictionary learning [11] is worse than PED [7] for Mode 1 and better than PED for mode 2. Our method outperforms both state-of-the-art methods – [11] and [7] for both modes.

V. CONCLUSION

In a previous work [11] it was shown that instead of using the standard Euclidean cost function for dictionary learning, better results are obtained with the more robust l_1-norm cost function. The reason has been discussed in the introduction. In this work, we extend the robust dictionary learning approach; we add a nuclear norm penalty on the coefficients. The reasoning behind this penalty is discussed in Section III.

The resulting formulation is solved using a combination of Split Bregman and Majorization Minimization. We compare our proposed technique with robust dictionary learning [11] and powerlet energy disaggregation [7]. Both are recent techniques, published within the last year. Our method outperforms both.

VI. ACKNOWLEDGEMENT

Authors acknowledge the support provided by ITRA project, funded by DEITy, Government of India, under grant with Ref. No. ITRA/15(57)/Mobile/HumanSense/01.

REFERENCES

[1] K. Carrie Armel, Abhay Gupta, Gireesh Shrimali and Adrian Albert, “Is disaggregation the holy grail of energy efficiency? The case of electricity”, Energy Policy, Vol. 52 (C), pp. 213-234, 2013.

[2] A. Gupta and P. Chakrabarty, “Impact of Energy Disaggregation on Consumer Behavior”, Behaviour, Energy and Climate Chance Conference, 2013.

[3] G. Crabtree, “Energy future report: “energy future: think efficiency”, American Physical Society, Tech. Rep., 2008.

[4] H. G.W., "Nonintrusive appliance load monitoring," Proceedings of the IEEE, Vol. 80, pp. 1870-1891, 1992.

[5] J. Z. Kolter and T. Jaakkola, “Approximate inference in additive factorial hmms with application to energy disaggregation,” in International Conference on Artificial Intelligence and Statistics, 2012, pp. 1472–1482.

[6] Z. Kolter, S. Batra, and A. Y. Ng., "Energy Disaggregation via Discriminative Sparse Coding," in Neural Information Processing Systems, 2010, pp. 1153-1161.

[7] E. Elhamifar and S. Sastry, “Energy Disaggregation via Learning 'Powerlets' and Sparse Coding”, AAAI 2015.

[8] A. Zoha, A. Gluhak, M. A. Imran and S. Rajasegarar, "Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey", Sensors, Vol.12, pp. 16838-16866, 2012.

[9] C. R. Paul, "Introduction to Electromagnetic Compatibility", Wiley, 2005.

[10] C. Fevotte and N. Dobigeon, "Nonlinear hyperspectral unmixing with robust nonnegative matrix factorization", IEEE Transactions on Image Processing, Vol. 24 (12), 2015.

[11] A. Majumdar and R. K. Ward, “Robust Dictionary Learning: Application to Signal Disaggregation”, IEEE ICASSP 2016 (accepted).

[12] P. J. Huber, “Robust Estimation of a Location Parameter”, The Annals of Mathematical Statistics, Vol. 35 (1), pp. 73-101, 1964.

[13] R. L. Branham Jr., "Alternatives to least squares", Astronomical Journal 87, pp. 928–937, 1982.

[14] M. Shi and M. A. Lukas, "An L1 estimation algorithm with degeneracy and linear constraints". Computational Statistics & Data Analysis, Vol. 39 (1), pp. 35–55, 2002.

[15] L. Wang, M. D. Gordon and J. Zhu, "Regularized Least Absolute Deviations Regression and an Efficient Algorithm for Parameter Tuning". IEEE ICDM. pp. 690–700, 2006.

[16] E. J. Schlossmacher, "An Iterative Technique for Absolute Deviations Curve Fitting". Journal of the American Statistical Association, Vol. 68 (344), pp. 857–859, 1973.

[17] G. O. Wesolowsky, "A new descent algorithm for the least absolute value regression problem". Communications in Statistics – Simulation and Computation, Vol. B10 (5), pp. 479–491, 1981.

[18] Y. Li and G. R. Arce, "A Maximum Likelihood Approach to Least Absolute Deviation Regression". EURASIP Journal on Applied Signal Processing, Vol. (12), pp. 1762–1769, 2004.

[19] A. Majumdar and R. K. Ward, “Some Empirical Advances in Matrix Completion”, Signal Processing, Vol. 91 (5), pp. 1334-1338, 2011.

[20] A. Majumdar and R. K. Ward, “Increasing Energy Efficiency in Sensor Networks: Blue Noise Sampling and Non-Convex Matrix Completion”, International Journal of Sensor Networks, Vol. 9, (3/4), pp. 158-169, 2011.

[21] http://cnx.org/contents/c9c730be-10b7-4d19-b1be-22f77682c902@3/Sparse_Signal_Restoration

[22] P. C. Hansen and D. P. O'Leary, “The use of the L-curve in the regularization of discrete ill-posed problems”, SIAM Journal on Scientific Computing, Vol. 14 (6), 1487-1503, 1993.

[23] J. Z. Kolter and M. J. Johnson , “REDD: A public data set for energy disaggregation research”, Proceedings of the SustKDD workshop on Data Mining Applications in Sustainability, 2012.