Handling Imperfections in Energy Disaggregation

Angshul Majumdar
IIIT Delhi
angshul@iiitd.ac.in

Megha Gupta
IIIT Delhi
meghag@iiitd.ac.in

Manoj Gulati
IIIT Delhi
manojg@iiitd.ac.in

***Abstract*— In this work we address two issues in energy disaggregation that are often overlooked – the problem of missing data, and, the problem of outliers. The first problem arises when the smart-meter cannot transmit the readings to the server owing to malfunctioning of the WiFi. The second problem arises from transients, surges and other non-linear effects. We modify the dictionary learning based disaggregation framework to address these problems. Results on the REDD dataset shows that results indeed improve when these issues are addressed explicitly.**

Keywords— Energy Disaggregation, Missing Data, Dictionary Learning, Robust Learning,

I. INTRODUCTION

In this work we propose to address two overlooked imperfections in energy disaggregation data. The first one is the problem of missing readings and the second one is regarding large yet sparse outliers.

In a typical experimental set-up, the data (training and testing) is collected by attaching smart-meters to individual appliances, which acquire the reading at periodic intervals and transmit these to a remote server wirelessly. Many a times the wireless connection does not work and hence the readings acquired during that period is not transmitted. This leads to missing data; it is common phenomenon in almost all energy disaggregation datasets. This is the first issue we propose to address in this work.

The second issue is regarding outliers. Energy disaggregation datasets are often corrupted by large but sparse noise (outlier). This may arise in several ways. It can occur from transients, voltage / current surges. In all such cases the anomaly is of large magnitude but for a very short duration. Such spikes are uncontrollable and do not show any characteristics of the appliance under study.

The other reason for occurrence of such spiky outliers 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 [1].
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 [2], 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.

The issues of data fidelity were first discussed in [3]; the author (of [3]) also released a code for pre-pre-processing such datasets [4]. Our work is different from the approach presented in [3]; we do not try to 'cure' the data by pre-processing, rather we mathematically model the imperfections.

The first problem of missing data is usually handled in a heuristic fashion by the NILM research community. The missing values either by imputed by prior readings or at best by nearest neighbour interpolation. In this work, we do not interpolate the missing readings, rather we modify the mathematical framework for disaggregation in order to accommodate information regarding missing readings.

The second problem, i.e. the problem of spiky outliers have been largely ignored by NILM researchers until recently [5]. Prior studies assumed that the noise in the system is small (Normally distributed) and hence employed a Euclidean data fidelity term. A recent study [5] argued about the existence of sparse but large outliers and following studies in robust estimation proposed a robust data fidelity term based on absolute deviations. In this work we follow the same approach.

II. LITERATURE REVIEW

The topic of Compressed Sensing (CS) has literally revolutionized signal processing research in the past decade. It studies the problem of solving an under-determined linear inverse problem where the solution is known to be sparse.

$$y_{m\times 1} = A_{m\times n}x_{n\times 1} + \eta_{m\times 1},\ \eta \sim \quad)\ m < n \qquad (1)$$

In general there are infinitely many solutions to (1). CS is interested in the case where the solution is s-sparse, i.e. x has only *s* non-zero values, the rest *n-s* being zeroes.

Donoho's seminal work [6] showed that a sparse solution is in most cases unique. CS literature shows that such a sparse solution can be recovered by l_1-minimization [6].

$$\min_x \|y - Ax\|_2^2 + \lambda \|x\|_1 \quad (2)$$

For most practical problems the signal is not sparse in itself but has an approximately sparse representation in the transform domain. Orthogonal and tight-frame transforms are useful in this case since the synthesis (3a) and analysis (3b) equations hold.

$$\text{synthesis: } x = W\alpha \quad (3a)$$

$$\text{analysis: } \alpha = W^T x \quad (3b)$$

where W is the sparsifying transform (wavelet / DCT for image, wavelet for EEG / ECG, STFT for speech etc.) and α the sparse transform coefficients.

This allows inverse problems arising from sparsifiable natural signals to be expressed in the following form,

$$y = AW\alpha + \eta \quad (4)$$

The sparse transform coefficients are solved using l_1-minimization (2) from which the signal of interest is obtained by applying the synthesis equation (3a). In CS lingo, A is called the measurement operator and W the sparsifying transform.

Dictionary learning gained popularity with the advent of K-SVD [7]. It was showed that instead of using fixed basis (like wavelet, DCT etc.) better solutions can be achieved by learning the sparsity basis from data. For dictionary learning, one needs training data (say X) from which a dictionary (D) is learnt such that the coefficients (Z) are sparse. The synthesis dictionary learning problem is framed as,

$$X = DZ \quad (5)$$

The learning can be framed in various ways. The basic constraints are Z should be sparse and there should not be degenerate solutions, i.e. very large D and small Z or vice versa. This can be prevented in several ways. K-SVD proposes an elegant (albeit slow) solution based on rank-1 updates. Others propose a normalization constraint on the columns of the dictionary. But the easiest formulation is to have a simple Frobenius norm penalty on the dictionary as a whole. This leads to the following formulation –

$$\min_{D,Z} \|X - DZ\|_F^2 + \lambda_1 \|D\|_F^2 + \lambda_1 \|Z\|_1 \quad (6)$$

This (6) constitutes the training phase. The learnt dictionary is used as the sparsifying basis in the testing phase for solving the sparse inverse problem.

Blind Compressed Sensing (BCS) [8] marries CS with dictionary learning. Instead of learning the dictionary in an offline fashion and then using it for solving the inverse problem, it learns the dictionary on the go. Obviously, it can be only used for multiple measurement vector (MMV) problems; this is because many samples would be needed to robustly estimate the dictionary.

Say the problem is to solve $Y = AX$; according to the dictionary learning formulation one assumes $X = DZ$; incorporating one into the other leads to –

$$Y = AX = ADZ \quad (7)$$

The solution to (7) is formulated as follows,

$$\min_{D,Z} \|X - ADZ\|_F^2 + \lambda_1 \|D\|_F^2 + \lambda_1 \|D\|_1 \quad (8)$$

III. Proposed Approach

A. Handling Missing Readings

The approach towards energy disaggregation is broadly based on the nature of the targeted household and commercial appliances. These appliances can be broadly categorised 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 [9]. 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) [10], 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 [11, 12] fall under this category.

Kolter et al [11], 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; this assumption is expressed in (9).

$$X_i = D_i Z_i,\ i = 1...\mathrm{N} \quad (9)$$

where D_i represents the codebook/dictionary and Z_i are the coefficients, assumed to be sparse. This is a typical dictionary learning problem with sparse coefficients.

As mentioned in the introduction, the training cannot be recorded owing to malfunctioning of the WiFi. Prior studies used to impute the missing data based on simple approaches such as nearest neighbour interpolation. In this work, we do not interpolate the data; rather we model the missing readings.

$$Y_i = R_i \odot \qquad \odot \quad (10)$$

where R_i is the binary sampling mask; it is 1 when the reading has been obtained and 0 when the reading is missing. Y_i is the data that is actually obtained (at the server).

The estimation problem (for dictionary and coefficients) is expressed as,

$$\min_{D_i,Z_i} \|Y_i - R_i \odot \qquad + \lambda_1 \|D_i\|_F^2 + \lambda_1 \|Z_i\|_1 \quad (11)$$

This is akin to the BCS formulation. The algorithm for solving this problem is available at [13].

The missing data problem exists also in the test / disaggregation phase. The aggregate reading is expressed as the power reading from individual appliances.

$$X = \sum_i X_i = [D_1 | ... | D_N] \begin{bmatrix} Z_1 \\ ... \\ Z_2 \end{bmatrix} \quad (12)$$

The dictionaries are obtained from training, the task at disaggregation is to estimate the coefficients Z_i's; this is done by simple l_1-minimization.

$$\min_Z \|X - DZ\|_F^2 + \lambda_2 \|Z\|_1 \quad (13)$$

where $D = [D_1 | ... | D_N]$ and $Z = \begin{bmatrix} Z_1 \\ ... \\ Z_2 \end{bmatrix}$.

Once the loading coefficients are estimated, the consumption for each appliance is obtained by:

$$\hat{X}_i = D_i Z_i, \; i = 1...\mathrm{N} \quad (14)$$

The issue of missing reading arises during disaggregation as well. A more appropriate model (compared to prior studies) would be express the problem as $Y = R \odot$. Thus the recovery is posed as a simple CS problem.

$$\min_Z \|Y - R \odot \qquad \lambda_2 \|Z\|_1 \quad (15)$$

B. Handling Outliers

In the previous sub-section we go by the assumption that the noise in the system, if any, is small. Hence the l_2-norm data fidelity term is justifiable. However, as discussed in the introduction, this is not the case. Training data is corrupted by sparse but large outliers arising from electrical surges and transients. The test data is corrupted not only by such surges and transients but also effects arising out of non-linear mixing. In such cases, where the noise appears as large and sparse outliers, the l_2-norm is not optimal.

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 [14] 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 [14]-[16]. The l_1-norm does not bloat the distance between the estimate and the outliers and hence is robust.

Following such studies in robust estimation, we propose to employ l1-norm data fidelity term in place of the l2-norm. For the scenario where the data is assumed to be complete (by simple interpolation) has been addressed in [3]. In this work we look at a more challenging problem – and model the missing readings. Therefore the corresponding formulation for the training (modifying (11)) and test / disaggregation (modifying (15)) phases are:

$$\min_{D_i, Z_i} \|Y_i - R_i \odot \qquad \cdot \lambda_1 \|D_i\|_F^2 + \lambda_1 \|Z_i\|_1 \quad (16)$$

$$\min_Z \|Y - R \odot \qquad \lambda_2 \|Z\|_1 \quad (17)$$

The solution for (17) is a standard one. Here we use the YALL1 algorithm [17]. However, solving (16) is not so straightforward; no off-the-shelf algorithm exists for (16). We derive an algorithm in the following sub-section.

C. Deriving a solution for (16)

Dropping the subscript 'i' from (16) (for the sake of simplicity) the task is to solve:

$$\min_{D,Z} \|Y - R \odot \qquad \lambda_1 \|D\|_F^2 + \lambda_1 \|Z\|_1$$

An elegant way to solve this problem is via the Split Bregman technique. We substitute $P = Y - R \odot$, and introduce the Bregman relaxation variable (B) leading to,

$$\min_{P,D,Z} \|P\|_1 + \lambda_1 \|D\|_F^2 + \lambda_1 \|Z\|_1 + \mu \|P - (Y - R \odot \qquad \}\|_F^2$$

This can be recast into the alternating minimization of the following sub-problems:

$$\mathrm{P1}: \min_P \|P\|_1 + \mu \|P - (Y - R \odot \qquad \}\|_F^2$$

$$\mathrm{P2}: \min_D \lambda_1 \|D\|_F^2 + \mu \|P - (Y - R \odot \qquad \}\|_F^2$$

$$\mathrm{P3}: \min_Z \lambda_1 \|Z\|_1 + \mu \|P - (Y - R \odot \qquad \}\|_F^2$$

P2 is the easiest to solve; it is a least squared problem having a closed form solution. P1 is has a closed form solution via soft thresholding. P3 needs to be solved iteratively via iterative soft thresholding. The usual constraint about positivity is enforced on the coefficient Z after every update.

The final step is to update the Bregman relaxation variable,

$$B \leftarrow P - (Y - R \odot \qquad \}$$

There are two stopping criteria for the Split Bregman algorithm. Iterations continue till the objective function converges (to a local minima) or if the maximum number of iterations reach 200.

IV. Experimental Results

We evaluate our proposed energy disaggregation framework on the REDD dataset [18]. The dataset consists of power consumption readings 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. The details of the dataset can be found in [18]. The 5th house is omitted since it does not have enough data. This evaluation protocol has been outlined in [18]. The evaluation metric has also been defined in the previous work as 'disaggregation accuracy'.

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

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 [3] it was shown that robust dictionary learning (with sparse coefficients) yields better results than standard the simple sparse coding proposed in [11] and the Factorial Hidden Model (FHMM) [10]. Therefore we compare our proposed approach only with [3]. We also compare ours with the powerlet based technique [12]. These [3, 12] are the most recent algorithms on energy disaggregation.

As outlined in [18], there are two modes of testing. In the Training mode, a portion of the data from every household is used as training samples and rest (from those households) is used for prediction. In the Testing mode, the data from four households are used for training and the remaining one is used for prediction; this is a more challenging problem.

Our proposed algorithm requires specifying two parameters (λ_1, λ_2) and one hyperparameter (μ). Some recent studies have shown that in a Split Bregman based technique, one can put the parameters to be unity and only tune the μ. We use the simple L-curve method to find out the hyperparameter; the value we obtained is μ=0.01.

Table. 1. Energy Disaggregation Results (in %)

House	Training Accuracy			Testing Accuracy		
	[3]	[12]	Prop	[3]	[12]	Prop
1	75.5	**81.6**	77.0	53.0	46.0	**54.5**
2	66.7	**79.0**	69.1	56.3	49.2	**58.0**
3	65.2	61.8	**67.0**	43.9	31.7	**45.7**
4	63.7	58.5	**65.9**	60.1	50.9	**61.6**
6	68.5	**79.1**	70.2	60.2	54.5	**62.0**

Comparison with [3] gives us insight regarding the importance of modelling missing readings. We show that instead of imputing the readings in a naïve fashion, better disaggregation results can be achieved if the missing values are modelled into the formulation.

We see that the powerlet based method [12] outperforms the simple yet robust learning techniques ([3] and proposed) for the case where the training data is scant but the test conditions are simple; training mode – data from the same house is for training and testing. But the powerlet based method shows poor performance when the training data volume increases but at the same time the problem becomes more challenging; testing mode – data from 4 houses are used for training and the 5th house is used for testing.

V. Conclusion

In this work we address two often ignored problems that are ever-present non-intrusive load monitoring. The first problem is of missing readings in the dataset – arising out of malfunctioning of the wireless network. The second problem is that of large and sparse outliers occurring out of transients, surges and non-linearities in the load.

The second problem has been handled in a recent work [5]. For the first time in this work, we are addressing the missing data and problem and combining it with the outlier removal problem in a single combined framework. It was shown in the prior study [5] that better results are indeed obtained when outliers are removed; in this work we show that the results can be further improved by accounting for the missing readings.

We have used the simplest possible formulation for dictionary learning. In [11] it was shown that better disaggregation results can be achieved when disaggregating terms are appended to the formulation in the learning phase. In the future we want to extend our proposed framework to such sophisticated learning formulations. We would also like to test our techniques on larger datasets.

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] M. Gulati, S. S. Ram, and A. Singh, "An In Depth Study into Using EMI Signatures for Appliance Identification", ACM BuildSys' 2014.

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

[3] S. Makonin, "Real-Time Embedded Low-Frequency Load Disaggregation", PhD Thesis, Simon Fraser University, 2014.

[4] https://github.com/smakonin/DataWrangle_REDD

[5] A. Majumdar and R. K. Ward, "Robust Dictionary Learning: Application to Signal Disaggregation", IEEE ICASSP 2016.

[6] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution", Comm. On Pure and Applied Maths, Vol. 59 (6), pp. 797 – 829, 2006.

[7] M. Aharon, M. Elad and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation", IEEE Transactions on Signal Processing, Vol. 54 (11), pp. 4311-4322, 2006.

[8] S. Gleichman, Y. C. Eldar, "Blind Compressed Sensing", IEEE Trans. on Information Theory 2011; 57: 6958-6975.

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

[10] J. Z. Kolter and T. Jaakkola, "Approximate inference in additive factorial hmms with application to energy disaggregation," AISTAT, 2012.

[11] Z. Kolter, S. Batra, and A. Y. Ng., "Energy Disaggregation via Discriminative Sparse Coding," NIPS, 2010.

[12] E. Elhamifar and S. Sastry, "Energy Disaggregation via Learning 'Powerlets' and Sparse Coding", AAAI, 2015.

[13] A. Majumdar, N. Ansari, H. Agarwal and P. Biyani, "Impulse Denoising for Hyper-Spectral Images: A Blind Compressed Sensing Approach", Signal Processing, Vol. 119, pp. 136-141, 2016.

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

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

[16] 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.

[17] http://yall1.blogs.rice.edu/

[18] J. Z. Kolter and M. J. Johnson , "REDD: A public data set for energy disaggregation research", SustKDD: workshop on Data Mining Applications in Sustainability, 2012.