Robust Dictionary Learning: Application to Signal Disaggregation

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

Rabab Ward
ECE, UBC
rababw@ece.ubc.ca

***Abstract*— It is well known that the Euclidean norm is sensitive to outliers; yet it is widely used for minimizing it is easy. Dictionary learning is no exception – the l_2-norm allows for easy update of the basis/dictionary atoms. In this work, we propose a robust dictionary learning method that is based on minimizing the robust l_1-norm. The ensuing optimization is solved using the Split Bregman approach. We apply the proposed technique to signal (energy and water) disaggregation and show that it excels over existing dictionary learning techniques (based on l_2-norm).**

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

I. Introduction

In signal disaggregation the task is to separate the aggregate signal into its individual components. Consider the case of energy disaggregation. One can only (in a non-intrusive fashion) record the total energy at the smart-meter; can we figure out what is the power consumption by individual appliances given the aggregate reading? The problem of water consumption disaggregation is similar. The total water consumption at regular instants of time is available to us; how can we find out the consumption of different sinks (e.g. cistern, tap, dishwasher, washer etc.)?

The motivation for disaggregating total energy and total water are somewhat different. In energy disaggregation, one is interested in knowing the consumption of individual appliances so that one can make an informed decision about using them. It has been observed that about 10-20% of power can be saved by changing user's behaviour [1] – [3]. In water consumption disaggregation the motivation is slightly different. When the consumption is known, it would have help identify leaks and other faults.

In both cases (energy and water) a linear mixing model is assumed, i.e. the aggregate signal is supposed to be a weighted sum of signals from individual components. Thus the task of disaggregation is to separate out the component-wise signals. There are several approaches to address this problem. The earliest approach in energy disaggregation was based on finite state machines [4]. More modern approaches generalize [4] to Hidden Markov Models [5]. The dictionary learning framework applies to both energy [6] and water consumption disaggregation [7].

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.

The l_2-norm minimization works when the deviations are small – approximately Normally distributed; but fail when there are large outliers. In statistics there is a large body of literature on robust estimation. The Huber function [8] 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 [9]-[11]. 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 [12]; Iterative Reweighted Least Squares [13] used to be another simple yet approximate technique. Other approaches include descent based method introduced by [14] and Maximum Likelihood approach [15].

In this work we propose robust dictionary learning, i.e. to learn the dictionary by minimizing the l_1-norm. The problem is solved using the Split Bregman technique. We apply our proposed method to the signal disaggregation problems and show that it performs better than the standard (non-robust) dictionary learning.

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

A. Disaggregation via Sparse Coding

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 basis, i.e. they expressed:

$$X_i = D_i Z_i,\ i = 1...\text{N} \tag{1}$$

where D_i represents the basis/dictionary and Z_i are the loading coefficients, assumed to be sparse. This is a typical dictionary learning problem with sparse coefficients. In [6] the dictionary learning problem is solved via:

$$\min_{D_i, Z_i} \left\| X_i - D_i Z_i \right\|_F^2 + \lambda \left\| Z_i \right\|_1 \quad (2)$$

Learning the basis 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 as:

$$X = \sum_i X_i \quad (3)$$

By imputing (1) in (3), one can express (3) as –

$$X = \left[D_1 | \ldots | D_N \right] \begin{bmatrix} Z_1 \\ \ldots \\ Z_N \end{bmatrix} \quad (4)$$

The loading coefficients can be solved using l_1-norm minimization.

$$\left\| X - \sum_i D_i Z_i \right\|_F^2 + \lambda \sum_i \left\| Z_i \right\|_1$$

(5)

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

$$\hat{X}_i = D_i Z_i, \; i = 1 \ldots \mathrm{N} \quad (6)$$

This approach for disaggregation was proposed for energy disaggregation initially [6]. Later it was shown that it could also be used for disaggregating water consumption [7]. In fact the linear mixing model assumed for energy disaggregation is not fully correct; it holds only for resistive loads but not for reactive loads – this is known from any book in first year electrical engineering. Most appliances consists of a composition of resistive and reactive loads.

The linear mixing model holds for water consumption. Here the total consumption is a sum of the consumption from individual sinks.

The model discussed here is the basic one. In both [6] and [7], other regularization terms were introduced to make the dictionary learning discriminative. This led to slight improvement in disaggregation results.

B. Dictionary Learning

In dictionary learning the goal is to learn an empirical basis from training data. The learnt basis may be used for a variety of tasks, e.g. inverse problems like denoising, reconstruction, etc. or for machine learning problems like classification and clustering.

One of the earliest known works in dictionary learning was based on the Method of Optimal Directions [16]. Given a training data X, they learnt a dictionary D and the codes Z by solving,

$$\min_{D,Z} \left\| X - DZ \right\|_F^2 \quad (7)$$

The learning consisted of alternately updating the dictionary / codebook D and the coefficients Z by coding.

Codebook update: $D_k \leftarrow \min_D \left\| X - DZ_{k-1} \right\|_F^2$ (8a)

Coding: $Z_k \leftarrow \min_Z \left\| X - D_k Z \right\|_F^2$ (8b)

In recent times, dictionary learning seeks to estimate a basis that can express the data in a sparse fashion. The problem is formulated as [17]:

$$\min_{D,Z} \left\| X - DZ \right\|_F^2 \; s.t. \; \left\| Z \right\|_0 \leq \tau \quad (9)$$

KSVD is a neat way to solve (9). It uses OMP for the sparse coding step and updates one column of the dictionary at a time using SVD.

However KSVD is slow, faster techniques for dictionary learning exists which are based on alternating minimization [18]. Such techniques directly solve the following,

$$\min_{D,Z} \left\| X - DZ \right\|_F^2 + \lambda \left\| Z \right\|_1 \quad (10)$$

Dictionary learning is a bilinear non-convex problem. There are some studies that prove convergence of KSVD type methods under some specific conditions; but in most cases these conditions are hard to satisfy.

III. Robust Dictionary Learning

The basic task of dictionary learning is to learn a basis given the training data. When X_i is the training data for i^{th} device, we express it as:

$$X_i = D_i Z_i \quad (11)$$

Prior studies in dictionary learning are based on minimizing an l_2-norm data mismatch – mainly because it is easy of minimization. The tacit assumption is that the mismatch follows a Normal distribution. In general one cannot make this assumption. If there are outliers the estimate from Euclidean norm minimization is skewed towards the outlier. We want a robust estimate. Therefore we propose to replace the l_2-norm by an l_1-norm data mismatch. When there is no requirement on the sparsity of the coefficients, this is expressed as follows:

$$\min_{D_i, Z_i} \left\| X_i - D_i Z_i \right\|_1 \quad (12)$$

Solving (12) may lead to a degenerate solution (this can happen to any alternating minimization based techniques –

this is not an issue with the l_1-norm). One may end up getting a very large value of *D* and very small values of *Z* so that the product remains finite; or the vice versa. In dictionary learning this problem is prevented by normalizing the columns of either *D* or *Z*.

When one needs sparse coefficients from the dictionary, an additional l_1-norm penalty on Z is to be added to (12).

$$\min_{D_i,Z_i} \left\| X_i - D_i Z_i \right\|_1 + \lambda \left\| Z_i \right\|_1 \tag{13}$$

Once the dictionary is learnt, we follow the procedure similar to [6]. The aggregate consumption (*X*) is assumed to the sum of consumptions from individual appliances (X_i's); we express –

$$X = \sum_i D_i Z_i \tag{14}$$

The individual loading coefficients are estimated by solving (15) – for dense coefficients and (16) for sparse coefficients.

$$\min_{Z_i} \left\| X - \sum_i D_i Z_i \right\|_1 \tag{15}$$

$$\min_{Z_i} \left\| X - \sum_i D_i Z_i \right\|_1 + \lambda \sum_i \left\| Z_i \right\|_1 \tag{16}$$

A. Deriving a solution for (12)

We introduce a proxy variable: *P=X-DZ*. The problem (12) is therefore expressed as,

$$\min_{D,Z,P} \left\| P \right\|_1 \; s.t.\, X = DZ \tag{17}$$

The unconstrained Lagrangian for (17) is,

$$L = \left\| P \right\|_1 + \mu^T \left(P - X + DZ \right) \tag{18}$$

The Lagrangian enforces strict equality; this is not required. One only needs to enforce strict equality at convergence. Therefore one can relax the equality constraint and use the Augmented Lagrangian instead.

$$AL = \left\| P \right\|_1 + \mu \left\| P - X + DZ \right\|_F^2 \tag{19}$$

The value of μ controls the relaxation; for small values the equality constraint between *P* and *X-DZ* is relaxed, and for high values it is enforced. One way to achieve this is to start with a small value of μ, solve (19); increase the value, solve (19) again and so on.

A more elegant solution is to introduce a Bregman relaxation variable (*B*) –

$$\min_{D,Z,P} \left\| P \right\|_1 + \mu \left\| P - X + DZ - B \right\|_F^2 \tag{20}$$

Instead of tinkering with μ, one can update *B* iteratively. The update is based on simple gradient descent and hence is very efficient. We only need to solve (20) once – for a fixed value of μ. Hence solving (20) is much less time consuming compared to (19). This approach is the so called Split Bregman technique.

One can segregate (20) into the alternating minimization of the following sub-problems:

$$\text{P1:} \min_{D} \left\| P - X + DZ - B \right\|_F^2 \tag{21a}$$

$$\text{P2:} \min_{Z} \left\| P - X + DZ - B \right\|_F^2$$

(21b)

$$\text{P3:} \min_{P} \left\| P \right\|_1 + \mu \left\| P - X + DZ - B \right\|_F^2 \tag{21c}$$

Solving P1 and P2 are straightforward – they are least squares problems and have closed form updates. They can also be solved using conjugate gradient based methods. Also P3 has a closed form update – soft thresholding [19].

B. Deriving a solution for (13)

To solve (13), we introduce the proxy variables as before and relaxing the equality constraint; this leads to:

$$\min_{D,Z,P} \left\| P \right\|_1 + \lambda \left\| Z \right\|_1 + \mu \left\| P - X + DZ - B \right\|_F^2 \tag{22}$$

Alternating minimization of (19) leads to the following sub-problems:

$$\text{P1:} \min_{D} \left\| P - X + DZ - B \right\|_F^2 \tag{23a}$$

$$\text{P2:} \min_{Z} \lambda \left\| Z \right\| + \mu \left\| P - X + DZ - B \right\|_F^2 \tag{23b}$$

$$\text{P3:} \min_{P} \left\| P \right\|_1 + \mu \left\| P - X + DZ - B \right\|_F^2 \tag{23c}$$

We have already discussed the solutions for P1 and P3. In this case, P2 does not have a single step update; instead it needs to be solved using Iterative Soft Thresholding [19].

C. Deriving a solution for (15)

This (15) is the standard l_1-norm minimization problem. Many solutions exists [12]-[15]. However, in this work we follow the Split Bregman approach we have been using so far to solve it. After introducing the proxy and relaxing the equality constraint we get,

$$\min_{Z,P} \left\| P \right\|_1 + \mu \left\| P - X + DZ - B \right\|_F^2 \tag{24}$$

In this case (24), alternating minimization would require solving sub-problems (21b) and (21c). We have already discussed the solution for these.

D. Deriving a solution for (16)

Following the Split Bregman approach we have been using so far, we have –

$$\min_{Z,P} \left\| P \right\|_1 + \lambda \left\| Z \right\|_1 + \mu \left\| P - X + DZ - B \right\|_F^2 \tag{25}$$

Alternating minimization of (25) would lead to two sub-problems (23b) and (23c). Techniques for solving them have already been discussed.

E. Updating the Bregman Relaxation Variable

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

$$B \leftarrow P - X + DZ - B \tag{26}$$

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

We evaluate our proposed signal disaggregation framework on energy disaggregation and water consumption disaggregation.

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 λ=1 and only tune the μ. We use the simple L-curve method [21].

A large-scale dataset contains nearly one million individual water use "events"; it was collected by Aquacraft from 1,188 residents in 12 study sites (such as Boulder, Colorado and Lompoc, California). For details, refer [20]. For this problem we use the F-measure as the evaluation metric [7]; it is defined as:

F-measure = 2 x precision x recall / (precision + recall)

Precision is the fraction of disaggregated consumption that is correctly classified while recall is the fraction of true device level consumption that is successfully separated

It is customary to report the evaluation metrics on both the training and the testing dataset. In the tables, the top value is the training error and the bottom value is the testing error. DDSC represents the standard dictionary learning technique [6], [7] where as FHMM denotes Factorial HMM.

Table. 1. Water Consumption Disaggregation

Device	DDSC	FHMM	Proposed – Dense Coeffn	Proposed – Sparse Coeffn
Toilet	56.44 40.46	53.51 49.70	55.32 50.16	59.98 50.15
Shower	72.37 30.97	58.29 53.27	70.19 56.04	75.76 56.11
Washer	49.13 22.23	21.19 25.97	45.52 30.27	55.30 30.29

*Training F-measure
Testing F-measure

For energy disaggregation, we report results on the REDD [22], [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. The disaggregation accuracy is defined as follows [22] –

$$Acc = 1 - \frac{\sum_{t}\sum_{n}\left|\hat{y}_t^{(i)} - y_t^{(i)}\right|}{2\sum_{t}\bar{y}_t} \tag{15}$$

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.

Table. 2. Energy Disaggregation Results (in %)

House	DDSC	FHMM	Proposed – Dense Coeffn	Proposed – Sparse Coeffn
1	73.3 50.0	71.5 46.6	70.1 52.1	75.5 53.0
2	63.7 55.7	59.6 50.8	61.9 55.7	66.7 56.3
3	62.1 40.8	59.6 33.3	61.0 43.2	65.2 43.9
4	70.9 55.6	69.0 52.0	71.0 59.8	73.7 60.1
6	65.4 58.9	62.9 55.7	64.7 60.0	68.5 60.2

*Training Accuracy
Testing Accuracy

In summary we see that robust dictionary learning indeed improves the disaggregation performance. Even with dense coefficients the testing performance is better than the standard dictionary learning (l_2-norm) DDSC and FHMM; the training performance is slightly less than DDSC but better than FHMM; however this is not an issue since improving the performance on testing is more challenging. Introducing sparsity in the learned coefficients improve the performance on training set but the improvement on testing set is nominal compared to robust dense dictionary learning.

V. Conclusion

In this work we propose a technique for robust dictionary learning; instead of minimizing the popular Euclidean norm cost function, we minimize the sum of absolute deviations – the l_1-norm. We test the performance on the problem of signal disaggregation and find that robust learning indeed improves performance for both water consumption disaggregation and energy disaggregation.

Our proposed method is unsupervised. There is a plethora of work in supervised dictionary learning for computer vision problems. In future we would like to extend our work robust dictionary learning to incorporate supervised dictionary learning penalties.

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] B. Wang, F. Chen, H. Dong, A. P. Boedihardjo and C. T. Lu, “Signal Disaggregation via Sparse Coding with Featured Discriminative Dictionary”, IEEE ICDM 2012, pp. 1134-1139.

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

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

[10] 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.

[11] 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.

[12] I. Barrodale and F. D. K. Roberts, "An improved algorithm for discrete L1 linear approximation". SIAM Journal on Numerical Analysis, Vol. 10 (5), pp. 839–848, 1973.

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

[14] 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.

[15] 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.

[16] K. Engan, S. O. Aase, J. Hakon Husoy, "Method of optimal directions for frame design," IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.5, pp.2443-2446, 1999

[17] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for Sparse Representation Modeling", Proceedings of the IEEE, Vol. 98 (6): 1045–1057, 2010.

[18] M. Yaghoobi, T. Blumensath and M. E. Davies, "Dictionary Learning for Sparse Approximations With the Majorization Method," IEEE Transactions on Signal Processing, Vol.57 (6), pp.2178-2191, 2009.

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

[20] P. Mayer, W. DeOreo, E. Opitz, J. Kiefer, W. Davis, and B. Dziegielewski, “Residential End Uses of Water," American Water Works Research Foundation, 1999.

[21] 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.

[22] http://redd.csail.mit.edu/

[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.