Robust Transform Learning

Jyoti Maggu
IIIT-Delhi
jyotim@iiitd.ac.in

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

***Abstract*— Dictionary learning follows a synthesis framework; the dictionary is learnt such that the data can be synthesized / re-generated from the coefficients. Transform learning on the other hand is based on analysis formulation; it learns a transform so as to generate the coefficients. The basic formulations of dictionary learning and transform learning employ a Euclidean cost function for the data fidelity term. Such cost functions are optimal when the noise / error in the system is Normally distributed, but not in the presence of sparse but large outliers. For such heavy tailed noise distributions, minimizing the absolute distance is more robust. There are several papers on robust dictionary learning. This work introduces robust transform learning. Experiments carried out on image analysis and impulse denoising elucidate the superiority of our method.**

Keywords*— *Analysis Dictionary Learning, Robust Estimation, Feature Extraction, Denoising

I. Introduction

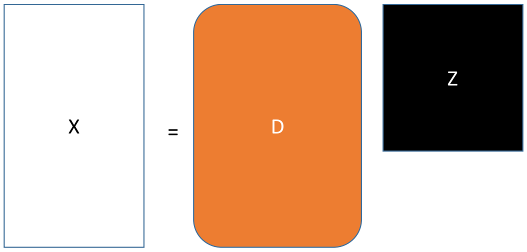

Fig. 1: Schematic Diagram for Dictionary Learning

In dictionary learning, we find a basis and learn the corresponding coefficients from the training data such that the basis / dictionary can synthesize / generate the training data. It was introduced in late 90's as an empirical tool to learn filters [1, 2]. The usual understanding of dictionary learning is shown in Fig. 1. The dictionary (*D*) and the coefficients (*Z*) are learnt from the data (*X*) such that the learnt dictionary and the coefficients can synthesize the data. Mathematically this is represented as,

$$X = DZ \tag{1}$$

Early studies in dictionary learning focused on learning a basis for representation. There were no constraints on the dictionary atoms or on the loading coefficients. The method of optimal directions [3] was used to learn the basis:

$$\min_{D,Z} \|X - DZ\|_F^2 \tag{2}$$

Here, *X* is the training data, *D* is the dictionary to be learnt and *Z* consists of the loading coefficients.

For problems in sparse representation, the objective is to learn a basis that can represent the samples in a sparse fashion, i.e. *Z* needs to be sparse. K-SVD [4] is the most well-known technique for solving this problem. Fundamentally, it solves a problem of the form:

$$\min_{D,Z} \|X - DZ\|_F^2 \text{ such that } \|Z\|_0 \le \tau \tag{3}$$

Here we have abused the notation slightly, the l_0-norm is defined on the vectorized version of Z. The problem with K-SVD is that it is slow, since it requires computing the SVD in every iterations and updating the coefficients via orthogonal matching pursuit. A faster solution to dictionary learning can be obtained by employing direct optimization methods [5].

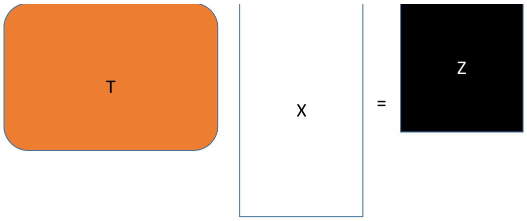

Fig. 2: Schematic Diagram for Transform Learning

Transform learning analyses the data by learning a transform / basis to produce coefficients. Mathematically this is expressed as,

$$TX = Z \tag{4}$$

Here *T* is the transform, *X* is the data and *Z* the corresponding coefficients. One may be enticed to solve the transform learning problem by formulating,

$$\min_{T,Z} \|TX - Z\|_F^2 + \mu \|Z\|_0 \tag{5}$$

Unfortunately such a formulation may lead to the trivial solution *T=0* and *Z=0*. In order to ameliorate this, the following formulation was proposed in [6] –

$$\min_{T,Z} \|TX - Z\|_F^2 + \lambda \left(\|T\|_F^2 - \log\det T \right) + \mu \|Z\|_0 \tag{6}$$

The factor $-\log\det T$ imposes a full rank on the learned transform; this prevents the degenerate solution. The additional penalty $\|T\|_F^2$ is to balance scale; without this $-\log\det T$ can keep on increasing producing degenerate results in the other extreme. Exactly the same formulation was restated as analysis dictionary learning in [7].

Dictionary learning has been employed extensively in solving various inverse problems [8-10]. Supervised dictionary has gained immense popularity in machine vision [11-13]. In all these studies, the data fidelity term is a Euclidean norm; this is based on the tacit assumption that the noise / error is Normally distributed. Even if that is not the case, researchers in signal processing and computer vision prefer employing the Euclidean norm owing to the relative ease of solution.

In signal processing literature, there have been studies where dictionary learning have been employed to solve problems that are known to be corrupted by sparse but large outliers. It has been used for solving impulse denoising problems [14, 15]; in recent times robust dictionary learning has been used in energy analytics where the signals are known to be corrupted by power surges which are sparse but of large magnitude [16, 17]. Owing to the heavy tailed distribution of such outliers, the aforementioned works [14-17] have proposed employing an l_1-norm for data fidelity cost.

In this work we propose a robust version of transform learning. This is achieved by replacing the l_2 / Frobenius norm of the data fidelity term in (6) by an l_1-norm. This makes the problem slightly difficult to solve. However following the Split Bregman technique we propose an efficient solution. We have carried out experiments on impulse denoising on images and have used it as a tool for feature extraction.

The rest of paper is organized in several sections. Since there are only a handful of studies in transform learning, we review it the following section. The proposed robust transform learning is described in section III. The experimental results are shown in section IV. The conclusions of this work are discussed in section V.

II. Transform Learning

We repeat the formulation for transform learning from (6).

$$\min_{T,Z} \|TX-Z\|_F^2 + \lambda\left(\varepsilon\|T\|_F^2 - \log\det T\right) + \mu\|Z\|_0$$

In [6, 18], an alternating minimization approach was proposed to solve the transform learning problem. This is given by –

$$Z \leftarrow \min_Z \|TX-Z\|_F^2 + \mu\|Z\|_0 \quad (7)$$

$$T \leftarrow \min_T \|TX-Z\|_F^2 + \lambda\left(\varepsilon\|T\|_F^2 - \log\det T\right) \quad (8)$$

Updating the coefficients (7) is straightforward. It can be updated via one step of Hard Thresholding. This is expressed as,

$$Z \leftarrow \left(abs(TX) \geq \mu\right) \odot TX \quad (9)$$

Here $\odot$ indicates element-wise product.

For updating the transform, one can notice that the gradients for different terms in (8) are easy to compute. Ignoring the constants this is given by –

$$\nabla\|TX-Z\|_F^2 = X^T\left(TX-Z\right)$$

$$\nabla\|T\|_F^2 = T$$

$$\nabla \log\det T = T^{-T}$$

In the initial paper on transform learning [6], a non-linear conjugate gradient based technique was proposed to solve the transform update. In the more refined version [18], with some linear algebraic tricks they were able to show that a closed form update exists for the transform.

$$XX^T + \lambda\varepsilon I = LL^T \quad (10)$$

$$L^{-1}XZ^T = USV^T \quad (11)$$

$$T = 0.5R\left(S + (S^2 + 2\lambda I)^{1/2}\right)Q^T L^{-1} \quad (12)$$

The first step is to compute the Cholesky decomposition; the decomposition exists since $XX^T + \lambda\varepsilon I$ is symmetric positive definite. The next step is to compute the full SVD. The final step is the update step. One must notice that L^{-1} is easy to compute since it is a lower triangular matrix.

The proof for convergence of such an update algorithm can be found in [19]. It was found that the transform learning was robust to initialization. There are only a few studies showing the application of transform learning; in [20] it was used for solving the MRI reconstruction problem.

III. Robust Transform Learning

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 [21] 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 [22]-[24]. 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 [25]; Iterative Reweighted Least Squares [26] used to be another simple yet approximate technique. Other approaches include descent based method introduced by [27] and Maximum Likelihood approach [28].

The IRLS have been before for solving the robust dictionary problem [14, 15]. The issue with IRLS is that, it solves a proxy of the actual problem; the convergence is only asymptotic. More recent studies on robust dictionary learning [16, 17] follow the Split Bregman approach. In this work we do the same.

For robust transform learning the formulation we propose is,

$$\min_{T,Z} \|TX-Z\|_1 + \lambda\left(\varepsilon\|T\|_F^2 - \log\det T\right) + \mu\|Z\|_0 \quad (13)$$

Here we have abused the notation slightly; the l_1-norm is defined on the vectorised data. The usage of the l_1-norm data fidelity term follows from the robust statistics.

We substitute $TX - Z = P$ in (13). This leads to the following constrained problem,

$$\min_{T,Z,P} \|P\|_1 + \lambda\left(\varepsilon\|T\|_F^2 - \log\det T\right) + \mu\|Z\|_0 \quad \text{such that } TX - Z = P \tag{14}$$

The Lagrangian for (14) is,

$$\min_{T,Z,P} \|P\|_1 + \lambda\left(\varepsilon\|T\|_F^2 - \log\det T\right) + \mu\|Z\|_0 + \eta\left(P - (TX - Z)\right) \tag{15}$$

The Lagrangian enforces equality in every iteration; this is not required in practice. One only needs to enforce equality at convergence. This can be achieved by the augmented Lagrangian,

$$\min_{T,Z,P} \|P\|_1 + \lambda\left(\varepsilon\|T\|_F^2 - \log\det T\right) + \mu\|Z\|_0 + \eta\|P - (TX - Z)\|_F^2 \tag{16}$$

The parameter η controls the equality constraint; a small value relaxes the constraint and a high value enforces it. One approach to solve the problem is to solve (16) for a small value of η and keep on heating (increasing) the value to enforce equality at convergence. However, this is not a particularly elegant approach. In this work, we follow the Split Bregman technique [29, 30]. We introduce a Bregman relaxation variable (B) which is updated in every iteration thereby automatically enforcing the equality at convergence.

$$\min_{T,Z,P} \|P\|_1 + \lambda\left(\varepsilon\|T\|_F^2 - \log\det T\right) + \mu\|Z\|_0 + \eta\|P - (TX - Z) - B\|_F^2 \tag{17}$$

One can segregate (17) into the alternating minimization of the following sub-problems:

$$\text{P1:} \min_{T} \|P - (TX - Z) - B\|_F^2 + \frac{\lambda}{\eta}\left(\varepsilon\|T\|_F^2 - \log\det T\right) \tag{18}$$

$$\text{P2:} \min_{Z} \|P - (TX - Z) - B\|_F^2 + \frac{\mu}{\eta}\|Z\|_0 \tag{19}$$

$$\text{P3:} \min_{P} \eta\|P - (TX - Z) - B\|_F^2 + \|P\|_1 \tag{20}$$

Solving for the Transform update (P1) follows from [18] after some algebraic manipulations.

$$XX^T + \lambda\varepsilon I = LL^T$$

$$L^{-1}X(P + Z - B)^T = USV^T$$

$$T = 0.5R\left(S + (S^2 + 2\lambda I)^{1/2}\right)Q^T L^{-1}$$

Updating the sparse coefficients Z from P2 is straightforward; requires one step of hard-thresholding.

$$Z \leftarrow \left(abs(TX - P + B) \geq \frac{\mu}{\eta}\right) \odot (TX - P + B)$$

Updating the proxy variable P from P3 is also straightforward, requiring one step of soft-thresholding.

$$P \leftarrow signum(TX + B - Z)\max\left(0, |TX + B - Z| - \frac{1}{\eta}\right)$$

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

$$B \leftarrow P - (TX - Z) - B$$

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 50.

IV. Experimental Results

We carry out two sets of experiments. In the first set, we employ robust transform learning to remove impulse noise from images. In the second set, we use robust transform learning as a feature extraction technique.

A. Impulse Noise Removal

In this work we compare the transform learning method with dictionary learning. We compare with the l_1-l_1 dictionary learning formulation for sparse noise removal in [14]. The dictionary learning based noise removal algorithm was formulated as –

$$\min_{D,Z,\hat{X}} \|X - \hat{X}\|_2^2 + \sum_i \|P_i\hat{X} - Dz_i\|_1 + \lambda\|z_i\|_1 \tag{21}$$

The first term is the data fidelity term between the noisy image X and the noisy version $\hat{X}$. The second term is the data fidelity term for dictionary learning; P_i is the operator that extracts patches from the image, D is the dictionary and z_i the sparse codes for the corresponding patch. The algorithm to solve it has been given in [14].

In this work we propose the transform learning version of image denoising. This has not been done before.

$$\min_{D,Z,\hat{X}} \|X - \hat{X}\|_2^2 + \sum_i \|TP_i\hat{X} - z_i\|_1 + \lambda\|z_i\|_0 \tag{22}$$

The first term is the global consistency term; this accounts for the removal of blocking artifacts arising from patch based operation. The second term arises from robust Transform learning; it is the l_1-norm since we are removing sparse noise. The solution to (22) can be obtained from alternate minimization.

$$\text{P1:} \min_{\text{T}} \sum_i \|TP_i\hat{X} - z_i\|_1 \tag{23a}$$

$$\text{P2:} \min_{\text{Z}} \sum_i \|TP_i\hat{X} - z_i\|_1 + \lambda\|z_i\|_0 \tag{23b}$$

$$\text{P3:}\min_{\hat{X}} \left\| X-\hat{X} \right\|_2^2 + \sum_i \left\| TP_i\hat{X}-z_i \right\|_1 \qquad (23c)$$

We have proposed algorithms to solve sub-problems P1 and P2. Sub-problem P3 is an l_1-minimization problem that can be solved using Iterative Soft Thresholding Algorithm [31].

We have carried out experiments on the Lena, Barbara and Baboon images. All of them were of size 256 x 256. These are standard images and we skip further description for the sake of brevity.

Fig. 3: Sample Test Images – Lena, Barbara and Baboon

The denoising results for several proportions of corruption (percentage of corrupted pixels) by salt-and-pepper noise are shown in the following table. The performance is measured in terms of structural similarity index (SSIM) which is known to be a better correlated with visual quality compared to PSNR.

Table. 1. SSIM Values after Denoising

Corrupted pixels	Lena		Barbara		Baboon	
	DL	TL	DL	TL	DL	TL
10%	0.97	0.98	0.95	0.95	0.96	0.97
30%	0.86	0.89	0.84	0.88	0.85	9.88
50%	0.78	0.82	0.77	0.82	0.78	0.83

*DL – Dictionary Learning [14]; TL – Proposed Transform Learning

The results clearly indicate that our proposed method yields superior results compared to dictionary learning based techniques for impulse noise removal. We have found that our results are robust to the values of λ between 0.01 and 1.

B. Feature Extraction

In [7] transform learning was dubbed as analysis sparse coding (ASC). They used it for feature extraction. In this work we compare robust transform learning with transform learning along with other representation learning tools – dictionary learning (K-SVD) [4], autoencoder (AE) and restricted Boltzmann machine (RBM). AE and RBM are used as basic building blocks for deep neural networks; however since transform learning and dictionary learning are shallow architectures, it is fair to compare with the basic AE and RBM. The results are shown in Tables 2 and 3 for Nearest Neighbour (NN) and Support Vector Machine (SVM) with rbf kernel.

We carried our experiments on several benchmark handwritten character recognition datasets. The first one is MNIST. It has 60K training samples and 10K test samples. The next one is USPS having 7291 training and 2007 test samples. Both MNIST and USPS are English digit datasets.

The Devnagari database [32], [33] of isolated handwritten Devnagari numerals consists of 22556 samples from 1049 persons. This database was formed from 368 mail pieces, 274 job application forms and for the rest we used a specially designed form for the present purpose.

The Bangla database [32], [33] of handwritten isolated Bangla numerals consists of 23392 samples written by 1106 persons. These samples had been collected from 465 mail pieces and 268 job application forms and for the rest, we used a specially designed form.

The number of nodes in AE and RBM are half the dimensionality of the input samples. The number of basis for KSVD, ASC/TL and proposed kernel Transform learning are one fourth the dimensionality of the input samples. These were found to yield the best results for all the datasets.

Table. 2. Classification Accuracy from NN

Dataset	AE	RBM	KSVD	ASC/TL	Proposed
MNIST	95.31	94.55	93.39	94.70	**96.59**
USPS	94.01	92.77	88.49	92.73	**94.02**
Devnagari	89.61	91.57	81.13	92.06	**92.14**
Bangla	78.80	**87.07**	77.60	79.17	85.07

Table. 3. Classification Accuracy from SVM

Dataset	AE	RBM	KSVD	ASC/TL	Proposed
MNIST	**96.62**	96.50	91.24	94.90	96.12
USPS	94.77	94.47	90.98	94.08	**94.77**
Devnagari	94.52	91.15	89.02	94.58	**96.25**
Bangla	90.90	89.37	85.28	91.87	**92.32**

We see that our proposed robust transform learning yields superior results not only to the basic transform learning / analysis sparse coding formulation but also compared to dictionary learning, RBM and AE on an average. This shows that robust transform learning has the potential to be used as a feature extraction tool.

V. CONCLUSION

Transform learning is an analysis basis learning framework. There are only a handful of studies on this problem. It has been used for image denoising and reconstruction. Under the name of analysis sparse coding, it has been used for feature extraction.

In this work, we propose robust transform learning. In the original formulation for transform / analysis sparse coding, the data fidelity term is a Euclidean norm. It is well known in robust statistics that such a norm is not robust towards outliers with heavy tailed distributions. In this work we replace the Euclidean norm by the sum of absolute distances. This makes our formulation less sensitive to outliers.

We have carried out experimental results on 1. inverse problem of impulse denoising; and 2. Feature extraction tool. In both cases we perform better. For impulse denoising it outperforms the l_1-l_1 dictionary learning and for feature extraction it yields better results than autoencoder, restricted Boltzmann machine, dictionary learning and transform learning.

REFERENCES

[1] Olshausen, B., and Field, D. 1997. Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Research. 37, 23, 3311-3325.

[2] Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. Nature. 401, 6755, 788-791.

[3] Engan, K., Aase, S., and Hakon-Husoy, J. 1999. Method of optimal directions for frame design. IEEE International Conference on Acoustics, Speech, and Signal Processing.

[4] Rubinstein, R., Bruckstein, A. M., and Elad, M. Dictionaries for Sparse Representation Modeling. Proceedings of the IEEE. 98, 6, 1045-1057.

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

[6] Ravishankar, S., and Bresler, Y. 2013. Learning Sparsifying Transforms. IEEE Transactions on Signal Processing. 61, 5, 1072-1086.

[7] Shekhar, S., Patel, V. M. and Chellappa, R. Analysis sparse coding models for image-based classification. IEEE International Conference on Image Processing. pp. 5207-5211, 2014.

[8] Protter, M., and Elad, M. 2009. Image Sequence Denoising via Sparse and Redundant Representations. IEEE Transactions on Image Processing. 18, 1, 27-35, 2009.

[9] Son, C.-H., and Choo, H. 2014. Local Learned Dictionaries Optimized to Edge Orientation for Inverse Halftoning. IEEE Transactions on Image Processing. 23, 6, 2542-2556.

[10] Caballero, J., Price, A. N., Rueckert, D., and Hajnal, J. V. 2014. Dictionary Learning and Time Sparsity for Dynamic MR Data Reconstruction. IEEE Transactions on Medical Imaging. 33, 4, 979-994.

[11] Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. 2008. Discriminative learned dictionaries for local image analysis. IEEE Conference of Computer Vision and Pattern Recognition.

[12] Yang, L., Jin, R., Sukthankar, R., and Jurie. F. 2008. Unifying discriminative visual codebook generation with classifier training for object category recognition. IEEE Conference of Computer Vision and Pattern Recognition.

[13] Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A., 2008. Supervised dictionary learning. Advances in Neural Information Processing Systems.

[14] Wang, S., Liu, Q., Xia, Y., Dong, P., Luo, J., Huang, Q., and Feng, D. D. 2013. Dictionary learning based impulse noise removal via L1–L1 minimization. Signal Processing, 93, 9, 2696-2708.

[15] Mukherjee, S., Basu, R., and C. S. Seelamantula, C. S. 2016. ℓ1-K-SVD: A robust dictionary learning algorithm with simultaneous update. Signal Processing. 123, 42-52.

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

[17] Gupta, M., and Majumdar, A. 2016. Nuclear Norm Regularized Robust Dictionary Learning for Energy Disaggregation. EUSIPCO.

[18] Ravishankar, S., Wen, B., and Bresler, Y. 2015. Online Sparsifying Transform Learning - Part I: Algorithms. IEEE Journal of Selected Topics in Signal Processing. 9, 4, 625-636.

[19] Ravishankar, S., and Bresler, Y. 2015. Online Sparsifying Transform Learning - Part II: Convergence Analysis. IEEE Journal of Selected Topics in Signal Processing. 9, 4, 637-746.

[20] Ravishankar, S., and Bresler, Y. 2015. Efficient Blind Compressed Sensing using Sparsifying Transforms with Convergence Guarantees and Application to MRI. SIAM Journal on Imaging Sciences, 8, 4, 2519-2557.

[21] Huber, P. J. 1964. Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35, 1, 73-101.

[22] Branham, R. L. 1982. Alternatives to least squares. Astronomical Journal, 87, 928-937.

[23] Shi, M., and Lukas, M. A. 2002. An L1 estimation algorithm with degeneracy and linear constraints. Computational Statistics & Data Analysis, 39, 1, 35-55.

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

[25] Barrodale, I., and Roberts, F. D. K. 1973. An improved algorithm for discrete L1 linear approximation. SIAM Journal on Numerical Analysis, 10, 5, 839-848.

[26] Schlossmacher, E. J. 1973. An Iterative Technique for Absolute Deviations Curve Fitting. Journal of the American Statistical Association, 68, 344, 857-859.

[27] Wesolowsky, G. O. 1981. A new descent algorithm for the least absolute value regression problem. Communications in Statistics - Simulation and Computation, B10, 5, 479-491.

[28] Li, Y., and Arce, G. R. 2004. A Maximum Likelihood Approach to Least Absolute Deviation Regression. EURASIP Journal on Applied Signal Processing, 12, 1762-1769.

[29] Goldstein, T., and Osher, S. 2009. The Split Bregman Method for L1 Regularized Problems. SIAM Journal on Imaging Sciences, 2, 2, 323—343.

[30] Nien, H., and Fessler, J. A.. 2014. A convergence proof of the split Bregman method for regularized least-squares problems. arXiv:1402.4371

[31] http://cnx.org/contents/c9c730be-10b7-4d19-b1be-22f77682c902@3/sparse-signal-restoration

[32] Bhattacharya U. and Chaudhuri, B. B. 2009. Handwritten numeral databases of Indian scripts and multistage recogni-tion of mixed numerals. IEEE Trans. Pattern Analysis and Machine Intelligence. 31, 3, 444-457.

[33] Bhattacharya U. and Chaudhuri, B. B. 2005. Databases for research on recognition of handwritten characters of Indian scripts. IEEE ICDAR.