Joint-sparse Dictionary Learning: Denoising Multiple Measurement Vectors

Prerna Singh
IIIT-Delhi
prerna13149@iiitd.ac.in

Ramy Hussein
UBC
ramy@ece.ubc.ca

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

Rabab Ward
UBC
rababw@ece.ubc.ca

***Abstract*— This work addresses the problem of denoising multiple measurement vectors having a common sparse support. Such problems arise in a variety of imaging problems, e.g. color imaging, multi-spectral and hyper-spectral imaging, multi-echo and multi-channel magnetic resonance imaging, etc. For such cases, denoising them piecemeal, one channel at a time, is not optimal; since it does not exploit the full structure (joint sparsity) of the problem. Joint-sparsity based methods have been used for solving such problems when the sparsifying transform is assumed to be fixed. In this work, we learn the sparsifying basis following the dictionary learning paradigm. Results on multi-spectral denoising and multi-echo MRI denoising demonstrates the superiority of our method over existing ones based on KSVD and BM4D.**

Keywords*— *denoising, dictionary learning, joint sparsity

I. INTRODUCTION

Mathematically additive noise is modelled as follows:

$$y = x + n \tag{1}$$

Here x is the clean image (to be recovered), n is the noise corrupting the image and y is the noisy obtained noisy image. The problem is to recover x given y and the noise characteristics of n.

This (1) is a single measurement vector problem. Denoising grey-scale digital images or X-ray images or MRI images fall under this category. However in a lot of problems, multiple structured or correlated measurements are available – these are multiple measurement vector (MMV) problems. For example in color imaging we have three channels – red, green and blue. These three channels are highly correlated with each other. Multi-spectral and hyper-spectral are extensions of color imaging, where the acquired bands stretch well beyond the visible range. Owing to the low spectral (band) gap (typically 5 to 10 nm for hyper-spectral images) images obtained at these bands are highly correlated with each other. Similar examples arise in multi-echo MRI [1] and multi-channel parallel MRI [2, 3].

All the aforesaid modalities are known to be corrupted by additive Gaussian noise. For such problems, the noise model for MMV can be expressed as follows:

$$Y = X + N \tag{2}$$

Here $Y = [y_1 | ... | y_M]$ and $X = [x_1 | ... | x_M]$ are multiple vectors. Moreover the multiple vectors X, to be recovered are correlated.

There are two ways to account for the correlation in the sparse recovery framework. One way is to 'whiten' or get rid of the correlation by some transform. This is the so called Kronecker Compressed Sensing formulation [5]. Here one transform sparsifies the vectors channel wise, and another transform whitens the inter-channel correlation. The other approach is to explicitly account for this correlation. It is assumed that owing to the correlation and structural similarity the positions of the non-zero values in the transform domain will remain same for all the vectors. This is the joint-sparsity assumption. This is the model of interest in this work.

Most traditional works in sparsity based denoising assumed the sparsity basis to be defined a priori, for example it can be the wavelet transform or the discrete cosine transform. Such transforms are generic and can sparsely represent varied class of signals; starting from digital photographs to medical images to speech and biomedical signals. Usually the denoising ability is limited by the sparsification capacity of these transforms; the sparser the representation, the better is the denoising performance. Such generic sparsifying transform, although applicable to a large class of signals, are not the best for any particular problem. Recent success of dictionary learning based techniques (such as [6]) empirically prove that.

In this work we learn the sparsity basis adaptively following the dictionary learning framework. We learn the dictionary such that it is capable of accounting for inter-channel correlations in the form of joint-sparsity. We will show that our method outperforms existing techniques for denoising multiple measurement vectors. Experiments have been carried out on multi-spectral imaging and multi-echo magnetic resonance imaging (MRI).

II. LITERATURE REVIEW

A. Kronecker Sparsity

Images are locally correlated. Sparsifying transforms like DCT and wavelet whitens the local correlation leading to a few high valued coefficients and the remaining coefficients are near about zero. This is why grey-scale or any single

channel image can be represented in a sparse fashion. Incorporating transform domain sparsity in (1) leads to:

$$y = S^T \alpha + n \tag{3}$$

where S is the sparsifying transform, assumed to be either orthogonal or tight framed.

For the single measurement problem, denoising can be framed as,

$$\min_{\alpha} \left\| y - S^T \alpha \right\|_2^2 + \lambda \left\| \alpha \right\|_1 \tag{4}$$

This is basically the classical wavelet thresholding technique to denoising [7]. Later extensions of this technique include [8, 9].

For multiple measurements, one needs to account for the fact that the multiple measurement vectors are correlated. In Kronecker sparsity based techniques, this correlation is whitened as well. For example in hyper-spectral images, the variation of pixel values in the spectral direction is gradual; similarly in dynamic CT imaging the temporal variation is smooth. Such variations can be effectively sparsified using Fourier transform (or temporal differencing).

Therefore in the multiple measurement vector model, one can apply a sparsifying transform S_1 in the spatial direction and another transform S_2 in the other (spectral / temporal) direction. The final sparse representation is $A=S_1XS_2$. Incorporating sparsity in both directions leads to,

$$Y = S_1^T A S_2^T + N \tag{5}$$

This can be conveniently represented as a Kronecker product $\otimes$ (hence the name).

$$vec(Y) = S_2 \otimes S_1^T vec(A) + vec(N) \tag{6}$$

The standard thresholding type techniques can be applied to solve (6), since (6) is mathematically equivalent to (3).

B. Joint-Sparsity

The Kronecker formulation REMOVES inter-channel correlation. The other approach is to preserve this correlation and effectively model it. For example in color imaging, since the red, green and blue channels are highly correlated, the image structure is the same in all the channels, i.e. the positions edges do not vary. In such a case, if wavelet transform is applied to each of the channels separately, the coefficients will be jointly sparse. This is because wavelet (and all other sparsifying transforms) effectively encodes the edge information; it has high values at the edges and low values elsewhere. Since the positions of the edges remain the same in all the channels, the positions of the non-zero values also remain the same in the transformed representation. Hence, when stacked as columns, they are joint-sparse or row-sparse.

One only needs to apply the spatial sparsifying transform to get $A=S_1X$. The coefficient matrix A will be row-sparse. The recovery is expressed as,

$$\min_{A} \left\| Y - S_1^T A \right\|_F^2 + \lambda \left\| A \right\|_{2,1} \tag{7}$$

The $l_{2,1}$-minimization problem for solving joint-sparsity problems has been known in signal processing for long [10]. Such models have been successfully used in color imaging [11] and MRI [3].

Both the Kronecker formulation and the joint-sparse formulation have their applicability and share of pros and cons. The Kronecker formulation can be used for capturing correlations that do not lead to joint-sparse structure, for example in dynamic medical imaging scenarios (MRI or CT). In such cases, the frames are the measurement vectors; since the content is changing, it does not lead to a block-sparse formulation, but can be well sparsified by applying a Fourier or finite difference [12, 13] transform in the temporal direction.

On the other hand for problems like parallel multi-channel MRI or multi-echo MRI Kronecker formulation is not applicable. This is because the variation across the multiple vectors does not follow a specific patter; therefore one does not know what kind of transform to apply for whitening the inter-channel correlations. In such cases, joint-sparsity is the way to go.

C. Dictionary Learning

So far we have discussed techniques where the sparsifying transform is given / fixed. They have certain advantages like fast operators, orthogonality etc. However sparsity based denoising techniques is heavily dependent on the sparsifying capacity of the transform – sparser the representation, better the denoising. Such fixed transforms do not yield the best possible representation for a particular class of signals. One needs to learn the sparsifying basis adaptively. This is achieved via dictionary learning.

One can learn the dictionary from generic image datasets and apply it for denoising a test image. However, there is no guarantee that such a basis (trained on a different set of images) will be the best representative for the test image. Therefore, in dictionary learning, one learns the basis from the image as it is denoising. For Gaussian denoising this is expressed as [6],

$$\min_{D,\mathrm{Z},\mathrm{x}} \left\| y - x \right\|_2^2 + \sum_i \left\| P_i x - D z_i \right\|_2^2 \text{ such that } \left\| z_i \right\|_0 \le \tau \tag{8}$$

The first term is for data fidelity. The term within the summation sign is for learning the dictionary and the coefficients. Here P_i represents the patch selection operator. It selects a patch (overlapping / non-overlapping) so that it can be sparse represented (z_i) by the dictionary D.

To the best of our knowledge there is a single comprehensive study on dictionary learning based color image denoising [14]. Basically it is the dictionary learning equivalent of the Kronecker formulation (6); however it learns a global dictionary from the patches of all the channels.

III. Proposed Formulation

In this work we propose the dictionary learning equivalent of the joint-sparse formulation. For single measurement vector problems, dictionary learning expresses the data as a sparse linear combination of its atoms.

$$x_i = Dz_i \tag{9}$$

Here x_i denotes the i^{th} patch.

For multiple measurement vector, the i^{th} patch will have components from different channels. In [14] a global dictionary is learnt for all the channels. This is expressed as

$$\begin{bmatrix} x_i^1 \\ \dots \\ x_i^M \end{bmatrix} = D \begin{bmatrix} z_i^1 \\ \dots \\ z_i^M \end{bmatrix} \tag{10}$$

Here we have assumed M channels. Some refinements were proposed in [14] over this base model, but it produced negligible improvement (less than 0.5 dB). Note that the dictionary learnt in (10) is significantly larger than the one in (9). For example if we take 8x8 patches, assuming a redundancy of 2, the dictionary in (9) is of size 64x128; but in (10) this is 192x384.

In this work we learn a single small dictionary, same size as (9), for all the channels. This is expressed as,

$$[x_i^1 \,|\, \dots \,|\, x_i^M] = D[z_i^1 \,|\, \dots \,|\, z_i^M] \tag{11}$$

But instead of imposing only sparsity as in (9), (10), we impose joint-sparsity, i.e. we assume that the rows of the coefficient matrix will be row-sparse. The reason behind this assumption has been discussed before. In short, it preserves the structure of the problem.

In order to address the color denoising problem, we incorporate group-sparsity into the grey-scale denoising formulation (8); this leads to,

$$\min_{D,Z,X} \|Y - X\|_2^2 + \sum_i \|P_i X - DZ_i\|_F^2 + \lambda \|Z_i\|_{2,1} \tag{12}$$

Here $P_i X = [x_i^1 \,|\, \dots \,|\, x_i^M]$, is the patch selection operator on all the channels; $Z_i = [z_i^1 \,|\, \dots \,|\, z_i^M]$. The $l_{2,1}$-norm enforces row-sparsity. Note that instead of using an NP hard penalty like l_0-minimization as in (8), we employ a convex surrogate $l_{2,1}$-norm.

This problem (12) has not been solved before. But is easy to solve using alternating minimization. This leads to the following sub-problems:

$$\text{P1}: \min_D \sum_i \|P_i X - DZ_i\|_F^2$$

$$\text{P2}: \min_X \|Y - X\|_2^2 + \sum_i \|P_i X - DZ_i\|_F^2$$

$$\text{P3}: \min_Z \sum_i \|P_i X - DZ_i\|_F^2 + \lambda \|Z_i\|_{2,1}$$

Sub-problems P1 and P2 are simple least squares minimization problems having an analytic solution in the form of pseudoinverse. Sub-problem P3 can be solved individually for each patch 'i'.

$$\min_{Z_i} \|P_i X - DZ_i\|_F^2 + \lambda \|Z_i\|_{2,1} \tag{13}$$

This can be solved efficiently using the modified iterative soft thresholding [15].

In this work we have used the standard redundancy of 2 in designing all the dictionaries. The dictionaries are initialized by concatenating an orthogonal Haar wavelet and a DCT transforms. This is a departure from prior dictionary learning techniques that uses a random subset of the patches for the same. The number of iterations (number of times the sub-problems are solved) is fixed at 20.

IV. Experimental Results

We carry out two sets of experiments. The first one is on multi-spectral image denoising and the second on multi-echo MRI denoising. In all the experiments, over-lapping patches of 8x8 were taken. The values in the overlapping regions were averaged. This applies to both our proposed techniques as well as KSVD [14]. Both of them use the standard redundancy of 2.

A. Multi-spectral Denoising

The experimental results were carried out on a well known CAVE multi-spectral imaging dataset [16]. The experiments were carried out on the four images - balloons, pompoms, cd_ms, clay_ms, chart and toys, feathers, glass tiles, sponges, flowers and bead. Experiments have been carried out on all 31 bands of each image.

Comparison has been carried out with the KSVD based technique proposed in [14]. In the aforesaid work [14], thorough experimentation had been carried out with other learning based methods like markov random fields [17]; it was found that the KSVD techniques yields significantly superior results. It is common knowledge that KSVD provides significantly better results than sparsity based techniques using fixed transforms; hence such techniques (e.g. wavelet denoising, TV denoising) are not compared here. We also compared against BM4D [18] – this is the state-of-the-art for denoising volumes.

The comparative results are shown in Table I. Owing to limitations in space, we are only showing SSIM (structural similarity index) [19] values; PSNR shows similar trend. Results are shown for Gaussian noise having σ=30 and σ=50.

TABLE I. SSIM For Different Techniques for 31 bands

Image	Proposed		KSVD [14]		BM4D [18]	
	σ=30	σ=50	σ=30	σ=50	σ=30	σ=50
balloons	**0.86**	**0.75**	0.82	0.67	0.84	0.70
pompoms	**0.86**	**0.76**	0.81	0.67	0.84	0.71
cd_ms	**0.86**	**0.73**	0.81	0.65	0.84	0.69
clay_ms	**0.77**	**0.60**	0.75	0.57	0.76	0.59
chart & toys	**0.75**	**0.59**	0.74	0.57	0.74	0.57
feathers	**0.80**	**0.66**	0.77	0.61	0.79	0.64

glass tiles	**0.82**	**0.68**	0.79	0.64	0.81	0.67
sponges	**0.87**	**0.74**	0.82	0.58	0.85	0.63
flowers	**0.79**	**0.58**	0.74	0.57	0.76	**0.58**
bead	0.71	**0.59**	**0.73**	0.58	0.71	**0.59**

The results show that except for the last image (where we are only marginally inferior), our method yields results superior to both the methods compared against. For visual clarity, the first four bands of the balloons image is shown in the Fig. 1. The images corroborate the numerical results. Visually, our proposed method gives almost perfect denoising, whereas KSVD and BM4D show severe artifacts.

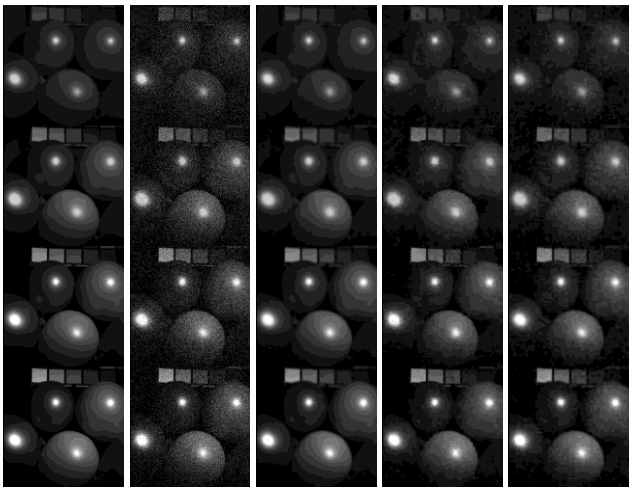

Fig. 1. Left to Right – Original , Noisy, Proposed, KSVD and BM4D

B. Multi-echo MRI Denoising

The experimental evaluation was carried out on ex-vivo and in-vivo T2 weighted images of rat's spinal cord. The data was collected with a Bruker 7T MRI scanner. The original data consisted of a series of 32 echoes acquired with a Carr-Purcell-Meiboom-Gill (CPMG) sequence with increasing echo time (first echo was acquired with 6.728 ms echo time, and consecutive echoes with the echo spacing of 6.738 ms). The data has been collected at the University of British Columbia.

For simulation studies, Gaussian noise (σ=10 and σ=20) is added in the Fourier domain (as is the physical principal in MRI) and the images are reconstructed by inverse Fourier transform. The output is complex since both the phase and the magnitude are preserved.

We compare our proposed technique with the same set of methods as before – KSVD [14] and BM4D [18]; both have been used before for denoising medical images. In these experiments, we report the SSIM from various techniques in Table II.

TABLE II. SSIM FOR DIFFERENT TECHNIQUES

Image	Proposed		KSVD [14]		BM4D [18]	
	σ=20	σ=30	σ=20	σ=30	σ=20	σ=30
In-vivo	**0.92**	**0.85**	0.87	0.81	0.89	0.84
Ex-vivo	**0.90**	**0.84**	0.84	0.80	0.89	0.81

The results corroborate the superiority of our method. For both the datasets, we give better results than KSVD and BM4D. For qualitative evaluation the results are shown in Fig. 2. In medical imaging it is customary to show both the recovered as well as the difference (between original and recovered) image. The difference image shows the recovery artifacts. These (one image each from the in-vivo and ex-vivo datasets) are shown in Fig. 2. It can be seen that the KSVD shows severe denoising artifacts; the artifacts are less pronounced in BM4D (but still visible in a good quality monitor). But with our proposed reconstruction, the artifacts are not palpable; the recovery is visually perfect.

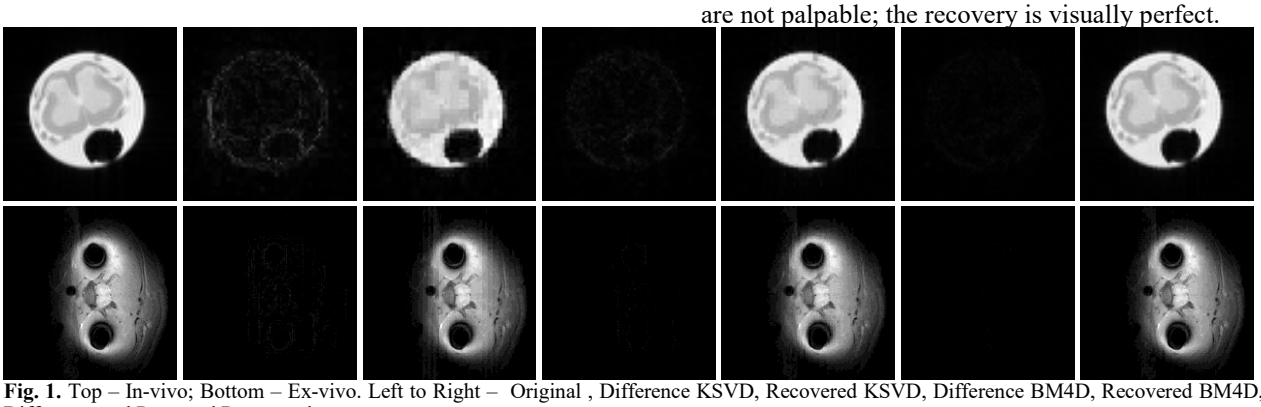

Fig. 1. Top – In-vivo; Bottom – Ex-vivo. Left to Right – Original , Difference KSVD, Recovered KSVD, Difference BM4D, Recovered BM4D, Proposed Difference and Proposed Recovered

V. CONCLUSION

This work proposes a new adaptive technique for removing noise from multiple measurement vectors (MMV). Comparison has been done with existing state-of-the-art – KSVD [14] for MMV and BM4D [18]. Our method excels over both.

However it must be noted that our method is only suitable for problems where the structure remains the same for all the measurements. It cannot handle volumetric data or time varying data; whereas the other methods [14, 18] can. But this is also the reason, why our method excels for multi-spectral and multi-echo denoising (owing to better modeling of structure) where the others do not.

ACKNOWLEDGEMENT

This work was made possible by NPRP grant 7-684-1-127 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] A. Majumdar and R. K. Ward, “Joint Reconstruction of Multi-echo MR Images Using Correlated Sparsity”, Magnetic Resonance Imaging, Vol. 29 (7), pp. 899-906, 2011.

[2] K. P. Pruessmann, “Encoding and reconstruction in parallel MRI”, NMR in Biomedicine, Vol.19(3), pp.288-299, 206.

[3] A. Majumdar and R. K. Ward, “Calibration-less Multi-Coil MR Image Reconstruction”, Magnetic Resonance Imaging, Vol. 30(7), pp. 1032-45, 2012.

[4] J. Fornaro, S. Leschka, D. Hibbeln, A. Butler, N. Anderson, G. Pache, H. Scheffel, S. Wildermuth, H. Alkadhi and P. Stolzmann, Dual-and multi-energy CT: approach to functional imaging. Insights into imaging, Vol. 2 (2), pp.149-159, 2011.

[5] M. F. Duarte and R. G. Baraniuk, "Kronecker Compressive Sensing", IEEE Transactions on Image Processing, Vol. 21 (2), pp. 494-504, 2012.

[6] M. Elad and M. Aharon, "Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries," IEEE Transactions on Image Processing, Vol.15 (12), pp. 3736-3745, 2006.

[7] D.L. Donoho, “De-noising by soft-thresholding”, IEEE transactions on information theory, Vol. 41(3), pp.613-627, 1995.

[8] S. G. Chang, B. Yu and M. Vetterli, “Adaptive wavelet thresholding for image denoising and compression”, IEEE transactions on image processing, Vol. 9 (9), pp.1532-1546, 2000.

[9] N. Weyrich and G. T. Warhola, "Wavelet shrinkage and generalized cross validation for image denoising", IEEE Transactions on Image Processing, Vol. 7 (1), pp. 82-90, 1998.

[10] Y. C. Eldar, P. Kuppinger and H. Bolcskei, “Block-Sparse Signals: Uncertainty Relations and Efficient Recovery”, IEEE Trans. Signal Processing, Vol. 58 (6), pp. 3042–3054, 2010.

[11] A. Majumdar, R. K. Ward and T. Aboulnasr, “Algorithms to Approximately Solve NP Hard Row-Sparse MMV Recovery Problem: Application to Compressive Color Imaging”, IEEE Journal on Emerging and Selected Topics in Circuits and Systems , Special Issue on Circuits, Systems and Algorithms for Compressive Sensing, Vol. 2 (3), pp. 362-369. 2013.

[12] G. Adluru, C. McGann, P. Speier, E. G. Kholmovski, A. Shaaban and E. V. R. DiBella, “Acquisition and Reconstruction of Undersampled Radial Data for Myocardial Perfusion Magnetic Resonance Imaging”, Journal of Magnetic Resonance Imaging, Vol. 29, pp. 466-473, 2009.

[13] L. Chen, M. C. Schabel, E. V. R. DiBella, “Reconstruction of dynamic contrast enhanced magnetic resonance imaging of the breast with temporal constraints”, Magnetic Resonance Imaging, Vol. 28 (5), pp. 637-645, 2010.

[14] J. Mairal, M. Elad and G. Sapiro, G, “Sparse representation for color image restoration”, IEEE Transactions on Image Processing, 17(1), pp.53-69, 2008.

[15] A. Majumdar and R. K. Ward, “Synthesis and Analysis Prior Algorithms for Joint-Sparse Recovery”, IEEE ICASSP, pp. 3421-3424, 2012.

[16] http://www.cs.columbia.edu/CAVE/databases/multispectral

[17] J. J. McAuley, T. S. Caetano, A. J. Smola and M. O. Franz, “Learning high-order mrf priors of color images”,. In ICML, pp. 617-624, 2006.

[18] M. Maggioni, V. Katkovnik, K. Egiazarian, A. Foi, "A Nonlocal Transform-Domain Filter for Volumetric Data Denoising and Reconstruction", IEEE Transactions on Image Processing, Vol. 22 (1), pp. 119-133, 2013.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity”, IEEE Transactions on Image Processing, Vol. 13 (4), pp. 600-612, 2014.