TV-DCT: Method to impute Gene Expression data using DCT based Sparsity and Total Variation Denoising

Akanksha Farswan, Anubha Gupta
{akankshaf,anubha}@iiitd.ac.in
Signal processing and Bio-medical Imaging Lab (SBILab)
Deptt. of ECE, IIIT-Delhi, New Delhi, India

Introduction

- High dimensional genomics data such as microarray gene expression data and RNA sequencing data, generally suffers from missing values.
- Incomplete data can adversely affect the downstream analysis for diagnostics and treatment.
- Several methods to impute missing values in gene expression data have been developed, but most of these work at high levels of observability.
- Interdependence between the expression levels of genes in gene expression data leads to a highly correlated data matrix (of subjects versus genes). Gene expression matrix can be considered as a low rank matrix embedded into a lower dimensional linear subspace. Further, missing value imputation in genomics can be viewed as a matrix completion problem.
- We propose a novel 2-stage method, namely, TV-DCT method for predicting missing values in gene expression data using Discrete Cosine Transform DCT based sparsity in Stage-1 and Total Variation (TV) denoising in Stage-2 .

Proposed Method

Proposed TV-DCT method for completing the gene expression matrix is a 2-stage method.

Stage-1: Compressive Sensing Framework

- Columns of gene expression matrix are highly sparse in the DCT domain because every column represents the expression values of a particular gene across subjects that would be biologically similar and hence, data within any column would be slowly varying in nature.
- Hence, we propose to recover missing data column-wise, i.e., by applying CS framework on each column of the matrix **Y**. The sensing matrix Φ_i of size $r_i \times m$ is constructed for every i^{th} column, where r_i denotes the number of observed entries in that column.
- Every i^{th} column of matrix **Y** using the CS-based reconstruction with the sparsity constraint on the column in the DCT domain as

$$\underset{\tilde{\mathbf{x}}_i}{min}(||\mathbf{y}_i - \Phi_i \tilde{\mathbf{x}}_i||_2^2 + \lambda_1 ||\mathbf{D}\tilde{\mathbf{x}}_i||_1) \quad (1)$$

- Since DCT is an orthogonal transform, we transformed it to synthesis prior formulation as

$$\underset{\tilde{\mathbf{z}}_i}{min}(||\mathbf{y}_i - \Phi_i \mathbf{D}^T \mathbf{z}_i||_2^2 + \lambda_1 ||\mathbf{z}_i||_1) \quad (2)$$

Stage-2: Denoising Framework

- Recovered matrix from stage-1 is assumed to be the noisy version of the original matrix and denoising is used in Stage-2. Total Variation (TV) based denoising is used in Stage-2 and is formulated as-

$$\underset{\mathbf{x}_i}{min}(||\mathbf{x}_i - \tilde{\mathbf{x}}_i||_2^2 + \lambda_2 ||\mathbf{A}\mathbf{x}_i||_1) \quad (3)$$

where i ranges from 1 to n (number of columns/ genes).

- **A** is a difference operator defined as

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \\ & & -1 & 1 \end{bmatrix}$$

- It maps a vector x_i to

$$(\mathbf{A}\mathbf{x}_i^k) = \mathbf{x}_i^k - \mathbf{x}_i^{k+1} \quad (4)$$

- Dual formulation of above is used to solve the optimization framework (because of non-differentiability of L_1-norm) as

$$\underset{\mathbf{x}_i}{min}\ \underset{|\mathbf{w}_i| \leq 1}{max}\ (||\mathbf{x}_i - \tilde{\mathbf{x}}_i||_2^2 + \lambda_2 \mathbf{w}_i^T \mathbf{A}\mathbf{x}_i) \quad (5)$$

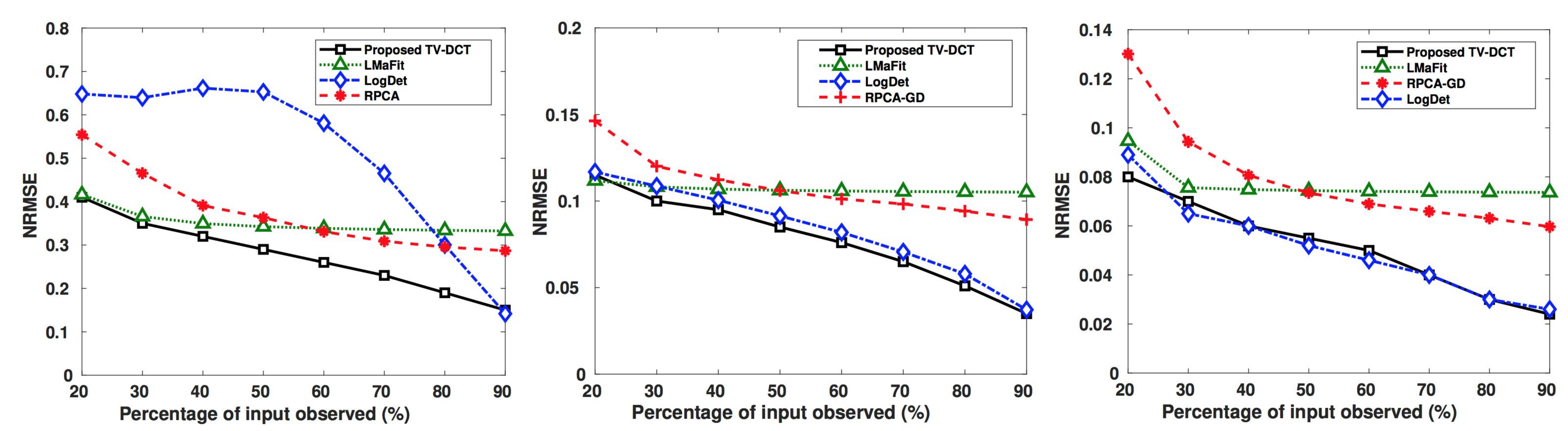
where w_i is an auxiliary vector such that

$$||\mathbf{x}_i||_1 = \underset{|\mathbf{w}_i| \leq 1}{max}\ (\mathbf{w}_i^T \mathbf{x}_i) \quad (6)$$

- TV denoising problem is minimized using iterative clipping algorithm with update equations as given in algorithm where,

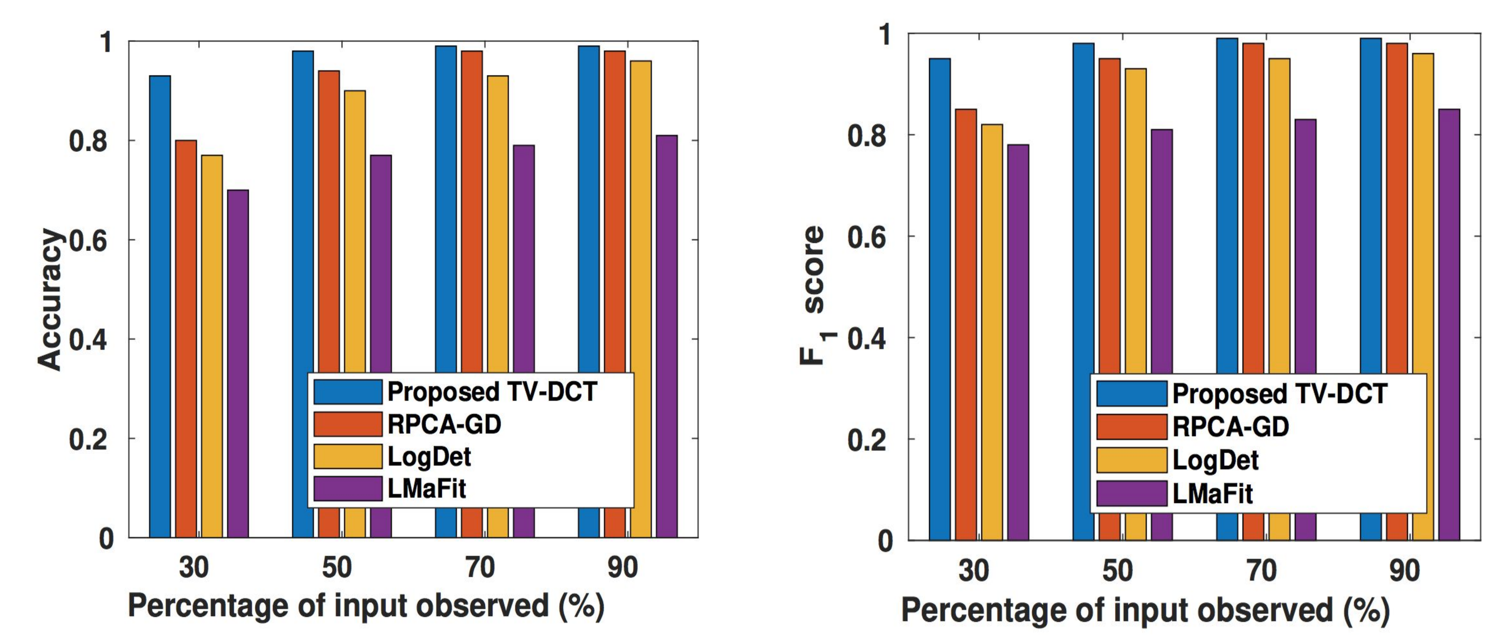
$$\mathbf{w}^{(0)} = \mathbf{0} \text{ and } \alpha \geq \text{maxeig}(\mathbf{A}\mathbf{A}^T)$$

Results



(a) ALLAML dataset (b) lung dataset (c) Myeloma dataset

Fig. 1: Comparison of proposed TV-DCT method with other methods at different percentage of observed input.



(a) Classification Accuracy (a) F_1 score

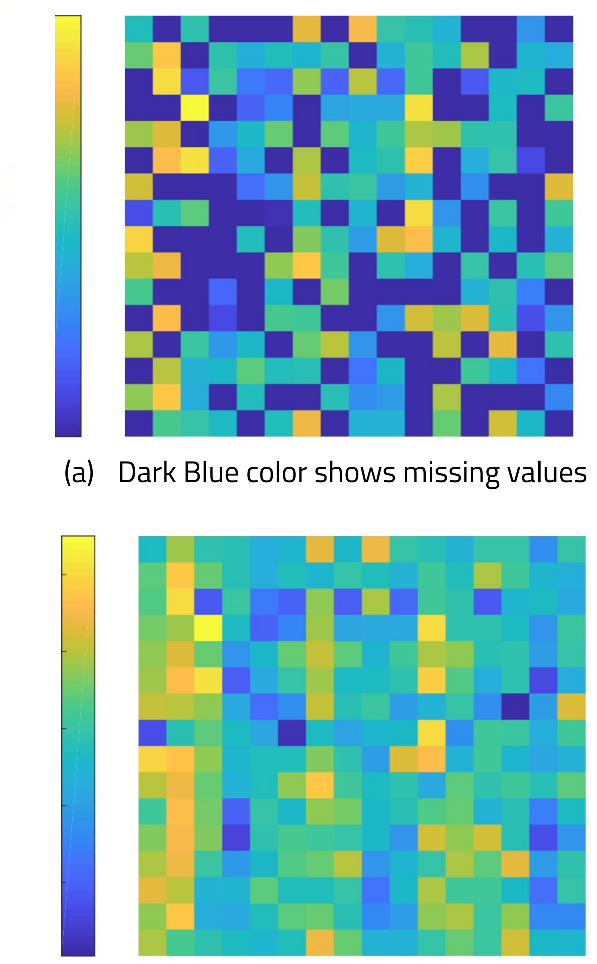
Fig. 2: Classification Accuracy and F_1 score obtained on imputed matrices of ALLAML dataset at varying sampling ratios.

Table 1:. Classification accuracy and F_1 scores on different sampling percentage of incomplete matrix and the recovered/imputed matrix using proposed TV-DCT method for ALLAML dataset

	Classification Accuracy				F_1 score			
Classifier	Random Forest		Linear SVM		Random Forest		Linear SVM	
SR	Observed	Imputed	Observed	Imputed	Observed	Imputed	Observed	Imputed
20	0.65	**0.96**	0.67	**0.90**	0.77	**0.96**	0.79	**0.90**
30	0.65	**0.96**	0.67	**0.93**	0.77	**0.96**	0.80	**0.95**
40	0.69	**0.97**	0.72	**0.94**	0.79	**0.97**	0.82	**0.95**
50	0.71	**0.97**	0.74	**0.97**	0.80	**0.97**	0.83	**0.98**
60	0.75	**0.97**	0.81	**0.98**	0.83	**0.97**	0.87	**0.99**
70	0.77	**0.96**	0.86	**0.99**	0.84	**0.97**	0.90	**0.99**
80	0.80	**0.95**	0.91	**0.99**	0.86	**0.96**	0.93	**0.99**
90	0.85	**0.95**	0.94	**0.99**	0.88	**0.96**	0.96	**0.99**

Algorithm

Stage 1 - Matrix Recovery
Input: **Y** (Input incomplete matrix), DCT matrix **D**
for loop from $i = 1........n$
Calculate Φ_i for all i using $\mathbf{y}_i$
while converge:
$\mathbf{z}_i^{k+1} = soft\left\{\mathbf{z}_i^k + \frac{1}{\alpha}(\mathbf{D}\Phi_i^T)(\mathbf{y}_i - \Phi_i \mathbf{D}^T \mathbf{z}_i^k), \frac{\lambda_1}{2\alpha}\right\}$
end while
$\tilde{\mathbf{x}}_i = \mathbf{D}^T \mathbf{z}_i$
end for
Obtain $\tilde{\mathbf{X}}$ from $\tilde{\mathbf{x}}_i$
Output: $\tilde{\mathbf{X}}$ (Recovered Matrix)
Stage 2 - Denoising
Input: $\tilde{\mathbf{X}}$(Noisy matrix), **A** (Difference Operator)
for loop from $i = 1........n$
while converge:
$\mathbf{x}_i^{k+1} = \tilde{\mathbf{x}}_i - \mathbf{A}^T \mathbf{w}_i^k$
$\mathbf{w}_i^{k+1} = clip\left\{\mathbf{w}_i^k + (\frac{1}{\alpha})\mathbf{A}\mathbf{x}_i^{k+1}, \frac{\lambda_2}{2}\right\}$
end while
end for
Obtain **X** from $\mathbf{x}_i$
$\mathbf{x}_{j,i} = \hat{\mathbf{x}}_{j,i}, \ \ if \ \ \Omega_{j,i} = 1$
Output: **X** (Recovered Matrix)



(a) Dark Blue color shows missing values

(b) Imputed Matrix

Fig. 1: 16x16 patch of Incomplete and Imputed matrix.

Evaluation metric is Normalized Root Mean Square Error defined as

$$\text{NRMSE} = \frac{||\hat{\mathbf{X}}(original) - \mathbf{X}(recovered)||_F}{||\hat{\mathbf{X}}(original)||_F} \quad (7)$$

Conclusion

- In this study, we have presented novel TV-DCT method that is a 2-stage matrix imputation method and we have investigated the performance of our proposed method at low as well as high observability of data.
- The comparative performance of TV-DCT method is observed to be superior to three state-of-the-art matrix completion methods in terms of NRMSE and classification accuracy.

References

1. J. Li et al., "Feature selection: A data perspective," ACM Computing Surveys (CSUR), vol. 50, no. 6, pp. 94, 2017.
2. Z. Wen et al., "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," Mathematical Programming Computation, vol. 4, no. 4, pp. 333–361, 2012.
3. Z. Kang et al., "Top-N recommender system via matrix completion.," in AAAI, 2016, pp. 179–185.
4. X. Yi et al., "Fast algorithms for robust pca via gradient descent," in Advances in NIPS, 2016, pp. 4152–4160
5. Ivan W Selesnick and Ilker Bayram, "Total variation filtering," White paper, 2010.
6. Leonid I Rudin, Stanley Osher, and Emad Fatemi, "Nonlinear total variation based noise removal algorithms," Physica D: nonlinear phenomena, vol. 60, no. 1-4, pp. 259–268, 1992
7. A. Gupta, S.D. Joshi, and P. Singh, "On the approximate discrete KLT of fractional Brownian motion and applications," Journal of the Franklin Institute, vol. 355, no. 17, pp. 8989–9016, 2018.
8. N. Jain, A. Gupta, and V. Ashok Bohara, "PCI-MDR: Missing Data Recovery in Wireless Sensor Networks using Partial Canonical Identity Matrix," IEEE Wireless Communications Letters, 2018.

Acknowledgements

Akanksha would like to thank UGC, Govt. of India for the UGC-JRF. We also acknowledge the funding support (Grant: EMR/2016/006183 and DST/ICPS/CPS-Individual/2018/279(G)) from DST, Govt. of India.