Sparse Representation of Cell Cycle Regulated Gene Expression in Yeast Using Cluster-Specific K-SVD Dictionary Learning

Yoon Jung*
400 Technology Square, Cambridge, MA 02139

ABSTRACT

A new method for achieving a sparse representation of gene expression data is proposed. Given a set of training data, PCA is performed for clustering analysis, followed by a K-SVD dictionary learning process for each cluster. Performance of data compression using the designed dictionary is evaluated using genome-wide expression data from DNA microarray hybridization.

Index Terms–K-SVD, dictionary learning, sparse coding, OMP, basis pursuit, FOCUSS, K-means, vector quantization, PCA

I. INTRODUCTION

Achieving a sparse representation of signals is an important task in many fields such as data compression, pattern recognition, and image processing [1-4]. Based on an overcomplete dictionary, expressing a signal as a linear combination of only a few atoms allows efficient storage and processing of data.

The sparsity of a signal highly depends on which domain it is expressed. For example, the wavelet transform-based JPEG 2000 shows superior performance to JPEG, since the original JPEG version uses the discrete cosine transform which requires more information to store an image file [5]. This means the image quality using JPEG 2000 will be better compared to the original JPEG version when the data compression rate is the same.

Although there are dictionaries such as wavelets that work well with various data types, they do not always guarantee a sparse representation. In some cases, the dictionary should be specific to the data type, being trained by a given set of signals. Here, I suggest a method which combines principal component analysis (PCA) and a dictionary learning algorithm named K-SVD, for sparse representation of gene expression profiles [6]. Data compression rate evaluation shows that K-SVD can store biological data efficiently, enabling further applications. Although not discussed here, the decomposed signal using the learned dictionary can be applied to logistic regression algorithms to classify genes to groups for diagnostic purposes [7]. Here, we simply focus on the data compression capability.

The contents in this article is as follows: Section II.1 explains the proposed method for compressing datasets that do not share common properties. In Section II.2, the singular value decomposition (SVD) algorithm and how it relates to principal component analysis (PCA) is explained. The flow of the K-SVD algorithm is discussed in Section II.3. To begin, the K-means algorithm and how it can be generalized into the K-SVD algorithm is briefly explained. Next, the sparse coding step using orthogonal matching pursuit (OMP), and how the atoms in the dictionary are updated using SVD are discussed in Section II.4. Finally in Section III, the proposed method is applied to yeast cell cycle gene expression data to show that K-SVD is capable of dimensionality reduction, and may be efficient for visualization and pattern recognition when combined with machine learning algorithms.

II. METHODS

Sparse representation of a signal highly depends on the dictionary, which can be designed using prespecified linear transforms or a set of training data. Here, we only consider methods that adapt dictionaries to training datasets, since they are based on fewer assumptions of the signal. In this case, there are two factors that determine the dictionary: the algorithm, and the training dataset.

Many dictionary learning algorithms have been proposed prior to K-SVD. Examples include maximum likelihood methods [8], method of optimal directions [9], and maximum a-posteriori probability approaches [10]. Among these approaches, K-SVD is one of the most popular dictionary learning algorithms. K-SVD has advantages in terms of flexibility to different pursuit algorithms in the sparse coding step, simple implementation, and low complexity giving fast convergence.

The training dataset is also an important factor that determines the dictionary. The training dataset and the new signals to be decomposed using the dictionary should share similar properties. Otherwise, the dictionary-transformation will only give a sparse representation for the training dataset, but not the new signals.

For example, consider a set of genes divided into groups based on similarity, evaluated by the correlation of gene expression profiles. A dictionary learned by datasets only from group 1 will not give a sparse representation in group 2. Moreover, A dictionary learned by datasets from all groups will not give sparse representations for new datasets. In both cases, dictionary-transforms will give sparsity when using the training datasets, but not for new signals.

* Email: yoonjung@mit.edu

II.1. Proposed algorithm using PCA and K-SVD

The proposed algorithm for compressing data with different properties is explained below. Consider a case where a small number of training datasets and a large amount of new measurements are given. The goal is to design a dictionary that compresses the new measurements.

- Perform PCA on the training dataset
- Visualize the training data using the first l principal components. l should be tuned accordingly to the visualized data
- Run the K-means clustering algorithm
- For each cluster, run the K-SVD algorithm to obtain a cluster-specific dictionary
- Visualize new measurements (data to be analyzed) on the previous scatter plot
- For each new measurement, use the dictionary specific to its group for decomposition

The underlying assumption of this method is that dictionary learning will perform better when training datasets have similar properties, rather than being sampled from various clusters.

II.2. PCA and its relation to SVD

PCA is a popular method for dimensionality reduction and visualization for clustering analysis. Given a set of vectors, it performs an orthogonal transformation into a set of linearly uncorrelated vectors. PCA is a powerful method since one can also use only the first l basis vectors to reduce the dimension, and extract common features of different data points.

We briefly mention its relation to SVD, a factorization of a matrix of the form

$$M = USV^H \tag{1}$$

$M \in \mathbb{R}^{m\times n}$ is the given matrix, which we assume to be real for simplicity. $U \in \mathbb{R}^{m\times m}$ and $V^H \in \mathbb{R}^{n\times n}$ are both unitary matrices, and $S \in \mathbb{R}^{m\times n}$ is a diagonal matrix. V^H denotes the conjugate transpose of V.

In PCA, the first principal component has the largest possible variance. The kth component is computed by subtracting the $k-1$ principal components from M. This turns out to be an orthogonal transformation given by

$$T = MW \tag{2}$$

where W is a unitary matrix whose columns are the right-singular vectors of M. Thus, PCA can be computed through SVD. The K-SVD algorithm also includes SVD for updating the dictionary during each iteration.

II.3. K-means algorithm and K-SVD

Consider a matrix $Y \in \mathbb{R}^{m\times n}$ where each column of Y represents a single data point in m-dimensional space. For example, Y could be a gene expression profile of n genes, measured at m timepoints. The goal of K-means clustering is to classify the data into clusters, where all data points in a cluster are represented by a single vector. Therefore, the vectors are divided into groups, giving the name vector quantization (VQ). The $L2$ norm is mostly used to minimize the distance between each data point and the vector that represents that cluster.

The sparse representation we discuss here using the learned dictionary allows each data to be a linear combination of several vectors. Therefore, we can consider K-SVD as a generalization of the K-means algorithm.

K-SVD alternates between sparse coding using the current dictionary and updating the atoms one by one. For the sparse coding step, obtaining the sparsest solution is an NP-hard problem. Alternatively, approximate solutions can be obtained using different constraints. Examples include basis pursuit which uses the $L1$ norm, focal underdetermined system solver (FOCUSS) which uses the Lp norm where $p \leq 1$. Here, we use a greedy algorithm called orthogonal matching pursuit (OMP), which gives suboptimal solutions and has low computational complexity.

III. RESULTS AND DISCUSSION

Cell cycle data from yeast was used to evaluate the performance of the algorithm. Due to the small size of the dataset, the number of datasets in each cluster after PCA was too small to train the dictionary. Therefore, the entire datasets were used in order to test whether sparse representation was possible with this data type.

Data were drawn during the cell division cycle after synchronization by alpha factor arrest [11] which gave 18 timepoints. Each dataset was given by a vector, where its entries represent the relative fluorescence intensity over a reference sample. Among 2467 genes, 2000 genes were used to train the dictionary and the remaining 467 genes were used to see whether the dictionary gave a sparse represenation. The K-SVD algorithm was implemented in MATLAB and executed on a i7-4790 Intel processor.

Using the training dataset, the average number of nonzero coefficients used in the linear combination of atoms was computed in order to test the validity of the K-SVD algorithm. The number of atoms in the dictionary was set to 500, and the average number of nonzero coefficients was 5.13. Next, we computed this number with the remaining data which was not used during the dictionary training process. For the remaining 467 genes, this was 6.16.

Although this number was much smaller than the number of atoms in the dictionary, the number of coefficients required to represent the data was only reduced by a factor of 3.

This might be due to the following reasons:

- The number of data points (18) was too small, while there might be a minimum number of coefficients required to represent the data
- Other ways of data transformation would have been required. The data was originally log-transformed, since each entry in the dataset vector represented the relative intensity of the fluorescent dyes (Cy5 / Cy3)
- The datasets used to train the data did not share similar properties, making it difficult to induce sparsity

As the number of coefficients required to represent the data only reduced by a factor of 3, this might suggest the necessity of the proposed algorithm, as mentioned in the last possible reason. If the training data can be classified into groups and for each group one runs the dictionary learning process, new measurements corresponding to a certain group will be sparse in the dictionary domain specific to that group. For this, gene expression profiles with larger sizes will be required in order to design an overcomplete dictionary for each group.

Using the proposed method, one can also combine with machine learning algorithms. A previous study on using K-SVD for breast cancer recurrence risk evaluation used the computed coefficients as inputs for a logistic classifier task [7]. Combining the proposed algorithm with this method will not only allow analyzing gene expression data but also enable more efficient ways for pattern recognition, data compression in general.

REFERENCES

1. D. L. Donoho, I. M. Johnstone, *Ideal denoising in an orthonormal basis chosen from a library of bases,* Comp. Rend. Acad. Sci., Vol. 319, pp. 1317–1322, 1994
2. R. Coifman, D. L. Donoho, *Translation invariant denoising, Wavelets and Statistics* New York: Springer–Verlag, Vol. 103, Lecture Notes in Statistics, pp. 120150, 1995
3. E. P. Simoncelli, W. T. Freeman, E. H. Adelsom, D. J. Heeger, *Shiftable multi-scale transforms,* IEEE Trans. Inf. Theory, Vol. 38, pp. 587–607, 1992
4. J. L. Starck, E. J. Candes, D. L. Donoho, *The curvelet transform for image denoising,* IEEE Trans. Image Process., Vol. 11, pp. 670–684, 2002
5. M. W. Marcellin, M. J. Gormish, A. Bilgin, M. P. Boliek, *An overview of JPEG-2000,* Proc. Data Compression Conf., pp. 523541, 2000
6. M. Aharon, M. Elad, A. Bruckstein, *K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation,* IEEE Trans. Image Process., Vol. 54, pp. 4311–4322, 2006
7. M. Mahrooghy, A. Ashraf, D. Dayea, *Sparse Representation of Multi Parametric DCE-MRI features using K-SVD for Classifying Gene Expression Based Breast Cancer Recurrence Risk,* Proc. of SPIE Vol. 9035
8. M. S. Lewicki, B. A. Olshausen, *A probabilistic framework for the adaptation and comparison of image codes,* J. Opt. Soc. Amer. A: Opt., Image Sci. Vision, Vol. 16, no. 7, pp. 1587–1601, 1999
9. K. Engan, S. O. Aase, J. H. Husy, *Multi-frame compression: Theory and design,* EURASIP Signal Process., Vol. 80, no. 10, pp. 2121–2140, 2000
10. K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, *Dictionary learning algorithms for sparse representation,* Neural Comp., Vol. 15, no. 2, pp. 349–396, 2003
11. M. Eisen, P. Spellman, P. O. Brown, *Cluster analysis and display of genome-wide expression patterns,* Proc. Natl. Acad. Sci., Vol. 95, pp. 14863–14868