Dictionary Learning for Sparse Representation and Classification of Neural Spikes

Ahmed H. Dallal, Yiran Chen, Douglas Weber, and Zhi-Hong Mao
Department of Electrical and Computer Engineering
Department of Bioengineering
University of Pittsburgh
Pittsburgh, PA 15261
ahd12@pitt.edu

Abstract—**Spike sorting is the problem of identifying and clustering neurons spiking activity from recorded extracellular electro-physiological data. This is important for experimental neuroscience. Existing approaches to solve this problem consist of three steps: spike detection, feature extraction, and clustering. In our method, we use Fisher discriminant based dictionary learning to learn dictionary, whose sub-dictionaries are class specific, and estimate discriminative sparse coding coefficients by minimizing the within class scatter and maximizing the between class scatter. Both the reconstruction error and coding coefficients are used for clustering the testing data. The dictionary learn the proper features specific to this problem. The proposed method has high reconstruction power and high clustering accuracy of testing data.**

I. INTRODUCTION

Analysis of neurons spiking activity, recorded under different behavioral and psychological conditions, is a crucial step in experimental neuroscience and its application [TODOREF 1],[TODOREF 2]. When a neuron, in the vicinity of an electrode, fires an action potential, the voltage recorded by this electrode is superimposed by a pattern waveform [TODOREF 3], [TODOREF 4]. Each neuron is assumed to have a consistent waveform and the shape of this waveform depends on the cells characteristics, coupling medium and relative position to electrode [TODOREF 4]. The existing acquisition systems enable us to record many channels simultaneously, however, the reliability of applications based on these recordings is highly affected by the accuracy of identifying and classifying individual neurons activity, i.e. spike sorting. Manual supervised spike sorting is possible [TODOREF 5], but, it cannot be done during the experiment run time, and is time and effort consuming, especially with multielectrode arrays. Hence, accurate computer aided detection and classification emerged.

Typically, the spike sorting algorithms consist of three main steps [TODOREF 3], [TODOREF 4]: (i) spike detection, (ii) extraction of discriminative features, and (iii) clustering of the spikes by the extracted features. Some popular algorithms use lower dimension representations of waveforms, e.g., principal component analysis (PCA) [TODOREF 4], [TODOREF 6] and wavelet coefficients [TODOREF 7] as a set of features for the subsequent clustering. Several techniques with varying complexity were used for clustering. These techniques include simple k-means clustering [TODOREF 4], Bayesian inference [TODOREF 8], superparamagnetic clustering of the wavelet coefficients[TODOREF 7], masked expectation maximization [TODOREF 9], matched filtering with discriminant analysis [TODOREF 10], and joint dictionary learning and mixture models [TODOREF 11]. However, many of the available algorithms require human intervention during the classification step, and they may fail to resolve overlapping spikes, or fail with missing data.

Sparse decomposition of neural signals can lead to interesting sorting results. In sparse decomposition, each signal is represented as a unique sparse combination of several dictionary elements or basis vectors [TODOREF 11]–[TODOREF 13]. Sparse representations are flexible in matching signals structures and robust to noise [TODOREF 13]. Carlson et al. [TODOREF 11] jointly used dictionary learning and clustering to analyze multiple spikes, where dictionary was used to get the sparse representation. On the other hand, Ekanadham et al. [TODOREF 12] used continuous basis pursuit for the sparse decomposition. Both [TODOREF 11] and [TODOREF 12] provided efficient approximation and promising solutions for the spike sorting problem

In this work, we propose the use of Fisher discriminant based dictionary learning (FDDL) [TODOREF 14] for decomposition and classification of detected spike waveforms. The space over which the signal could be decomposed is learned from training samples. The learned structured dictionary is composed of class specific sub-dictionaries. To ensure discriminative coefficients, Fisher discrimination criterion is imposed when learning the dictionary and coding coefficients. The learned sparse coding coefficients and reconstruction errors are both used for classification of testing data sets.

The remainder of this paper is organized as follows. Section II describes the dictionary learning process, clustering criterion, and experimental testing setup. Section III begins with discussing the spike reconstruction capability of the learned dictionary, and then compares the performance of our approach with other existing clustering algorithms with highlighting additional features of our methods. Summary and future direction are given in Section IV.

II. Methods

In sparse representation classification problems, there are two phases: coding and classification. The dictionary atoms are first learned using training data samples. Then, the detected spiked are coded over this learned dictionary. The coding sparse coefficients are used along with reconstruction error to classify the detected spikes.

A. Spike Detection

First, the signal is filtered by a fourth order bandpass (300 – 6000 Hz) filter. Then, amplitude thresholding was used for spike detection. The threshold, Tv, is derived from an estimate of the noise level [TODOREF 7] as,

$$T_v = 4\hat{\sigma}_n, \quad \hat{\omega}_n = \text{median}(V(t)/0.6745), \quad (1)$$

where V (t) is the filtered signal and TODO is an estimate of noise standard deviation. For each detected spike, a 2.5 milliseconds widow is identified as the spiking activity to be processed.

B. Fisher Discrimination Dictionary Learning

A structured dictionary, D, is learned based on a Fisher iscriminant scheme [TODOREF 14], such that the resultant dictionary atoms have correspondence to class labels, i.e., $D = D_1, D_2, ..., D_N]$, where D_i s the class specific sub-dictionary sociated with class i, and N is the total number of classes. or a training samples set $A = [A_1, A_2, ..., A_N]$, where A_i s he sub-set samples from class i, we denote X as the coding oefficients of samples matrix A over D, i.e., $A \approx DX$. Also, can be expressed as $X = [X_1, X_2, ..., X_N]$, where X_i s he sub-matrix for coding coefficients of A_i over D.

The coding coefficients should be sparse, and the dictionary should be able to accurately construct A. The dictionary should have a high discriminating capability of spike classes in A, as well. Based on these three objective, the optimization problem in (TODOREF 2) can be proposed to estimate D and X from the training set A.

$$J = \min_{D,X} \sum_{i=0}^{N} r(A_i, D, X_i) + \lambda\|X\|_1 + \lambda_2 f X \quad (2)$$

where $r(A, D, X)$ is the discrimination fidelity term, $\|X\|_1$ is the sparsity constraint, $f(X)$ is the discrimination constraint on the coding coefficients, and $\lambda_{1,2}$ are scalar parameters.

With each sub-dictionary D_i associated with the ith class, A_i should be well represented by D_i but not D_j, $i \neq j$. This implies that X_i be nearly zero, except for some significant coefficients that correspond to the sub-dictionary D_i. Hence, the discriminative fidelity term is defined as

$$r(A_i, D, X_i) = \|A_i - DX_i\|_F^2 + \|A_i - D_i X_i^i\| +$$
$$\sum_{j=1,j\neq i}^{N} \left\| D_j X_i^j \right\|_F^2, \quad (3)$$

where X_j^i is the coding coefficient of A_i over D_j. Discriminative coding coefficients is achieved by designing $f(X)$ using

Fisher Discrimination Dictionary Learning
1. Initialize atoms of D with random vectors of unit norm.
2. Fix D and update the sparse coding coefficients $X_i, i = 1, \ldots, N$, one by one; by solving (5).
3. Fix X and update each sub-dictionary $D_i, i = 1, \ldots, N$, by solving (6).
4. Repeat steps 2 and 3, until $J \leq \epsilon$ or the maximum number of iterations is reached.

Fisher discrimination criterion. This criterion minimizes the within class variance of X, $S_W(X) = \sum_{i=1}^{N} \sum_{x \in X_i} (\mathbf{x} - \mathbf{m_i})(\mathbf{x} - \mathbf{m_i})^T$, and maximizes the between class variance of X, $\sum_{i=1}^{N} n_i(\mathbf{m} - \mathbf{m_i})(\mathbf{m} - \mathbf{m_i})^T$; where n_i is the number of training samples in class i, and $\mathbf{m}$ and $\mathbf{m_i}$ are the mean vectors of X and X_i respectively. Therefore, a convex discrimination function could be defined as

$$f(X) = \text{tr}(S_W(X)) - \text{tr}(S_B(X)) + \eta\|X\|_F^2, \quad (4)$$

where the last term is an elastic term with a scalar parameter η. Substituting (TODOREF 3) and (TODOREF 4) into (TODOREF 2) completes the problem formulation. The objective function (TODOREF 2) is not jointly convex on D and X, but, it is convex with respect to each of them when the other is fixed. Therefore, D and X are alternatively optimized, where we update X by fixing D, and update D by fixing X. These two sub-problems are iteratively solved until convergence.

In the first sub-problem, (TODOREF 2) is reduced to sparse coding problem in which Xi was estimated class by class. When estimating Xi , other classes coefficients Xj , j 6= i, are held constant. The reduced objective function is

$$J_{X_i} = \min_{X_i} r(A_i, D, X_i) + \lambda_1\|X\|_1 + \lambda_2 f_i(X_i), \quad (5)$$

with $f_i(X_i) = \|X_i - M_i\|_F^2 - \sum_{k=1}^{N} \|M - M_k\|_F^2 + \|X\|_F^2$, where M_k and and M are the mean vector matrices of class k and all classes receptively. We set η to 1, so that $f_i(X_i)$ is strictly convex to X_i. Equation (5) is solved using the iterative projection method [TODOREF 14].

The sub-dictionaries D_i are estimated class by class, while fixing X and all other sub-dictionaries D_j, $j \neq i$. The reduced objective function is

$$J_{D_i} = \min_{D_i} \left\| A - D_i X^i - \sum_{j=1,j\neq i}^{N} D_j X_i^i \right\|_F^2 +$$
$$\|A_i D_i X_i^i\|_F^2 + \sum_{j=1,j\neq i}^{N} \|D_i X_j^i\|_F^2, \quad (6)$$

which is a quadratic programming problem. The algorithm for FDDL is shown in table TABLEREF.

C. Spike Clustering

Testing samples are classified by coding them over D. Each sample, s, is coded over each class specific sub-dictionary, and

then the reconstruction error along with the coding coefficients are used for classification.

Since the number of training samples is relatively with respect to the number of clusters, D_i can well represent the ith class. Hence, the test sample, s, is directly codded over D_i; this reduces the influence of other classes sub-dictionaries. The coding coefficients, y_i, associated with D_i can then computed as

$$\hat{\mathbf{y}}_i = \operatorname*{argmin}_{\mathbf{y}_i} \|\mathbf{s} - D_i \mathbf{y}_i\|_2^2 + \gamma_1 \|\mathbf{y}_i\|_1 + \gamma_2 \|\mathbf{y}_i - \mathbf{m}_i^i\|_2^2, \quad (7)$$

where TODOFIX$m_i{}^i$ is the ith-class trained mean sub-vector associated with D_i, and γ_1 and γ_2 are constants. From (TODOREF 7), y_i should be close to TODOFIX$m_i i$, and D_i should well sparsely code s. Therefore, the classification metric is defined by

$$e_i = \|\mathbf{s} - D_i \hat{\mathbf{y}}_i\|_2^2 + \gamma_1 \|\hat{\mathbf{y}}_i\|_1 + \gamma_2 \|\mathbf{y}_i \hat{-} \mathbf{m}_i^i\|_2^2, \quad (8)$$

The testing sample s was assigned to the class that gave the minimum error (TODOREF 9).

$$\operatorname{cluster}(\mathbf{s}) = \operatorname*{argmin}_i e_i \quad (9)$$

D. Testing Experiment

The algorithm was applied on four simulated data sets [TODOREF 7], each contains spiking activities from three different neurons with mean firing rate of 20 Hz and 24 KHz sampling rate. The data were taken from real recordings, and four noise levels were imposed to simulate real background activity. The noise level was from the standard deviation, which was equal to 0.05, 0.1, 0.15, and 0.2 relative to the spike classes amplitude. Fig. 1 shows a sample of the used data sets. The reason of using simulated data is to obtain an objective measure of performance, since the simulation defines the identity of the spikes.

The dictionary learning and clustering parameters $(\lambda 1, \lambda 2, \gamma 1, \gamma 2)$ were evaluated by 5-fold cross validation. The chosen values for these parameters are $\lambda 1 = \gamma 1 = 0.005$ and $\lambda 2 = \gamma 2 = 0.05$.

We compared our method to the superparamagnetic clustering (SPC), introduced in [TODOREF 7], and to the standard k-means clustering [TODOREF 4], where k-means clustering was applied to the PCA components of the detected spikes. Here, we chose the components that retained 99% of data variance.

III. RESULTS AND DISCUSSION

The dictionary learning was done with 100 training sample of each spike class, approximately 10% of the simulation data set. The construction capability is one powerful aspect of using dictionary learning, and this was emphasized in the objective function used to build learn the dictionary (TODOREF 2). A reconstructed spike samples are shown in Fig. 2, it is shown that the spikes were well reconstructed, with a mean squared error (MSE) of 0.015 and 0.017, respectively. The total reconstruction mean squared error for the training set was 3.49.

TABLE II
ACCURACY OF CLUSTERING TESTING SPIKING SAMPLES (MEAN ± STD)

Method	Noise Level			
	0.05	0.1	0.15	0.2
FDDL	0.97 ± 0.05	0.88 ± 0.08	0.76 ± 0.06	0.66 ± 0.10
SPC	0.65 ± 0.01	0.61 ± 0.29	0.37 ± 0.03	0.19 ± 0.11
k-means	0.50 ± 0.36	0.48 ± 0.27	0.45 ± 0.38	0.40 ± 0.34

These results show the reconstruction power and accuracy of the learned dictionary. The fidelity term in (TODOREF 2) optimized for the reconstruction error by each class specific sub-dictionary, this along with the discrimination constraint, $f(x)$, which maximized the between classes variance, have reduced the interference from other classes, that is reflected as low variance of the reconstructed signal from the original one. This powerful capability enables us to use sparse representation of signals without a sever loss of data; besides that, this sparse representation can be also used for classification. This could also enable estimation of missing data points or clipped signals.

The accuracy of correctly classifying testing detected spikes waveform was calculated for three methods: FDDL dictionary learning, standard clustering using PCA and and k-means [TODOREF 4], and the superparamagnetic clustering [TODOREF 7]. Table II shows the accuracy of the three methods under different noise levels, the average was taken for ten experiments and four

data sets. During each experiment the dictionary was learned with randomly picked 10% of each class data, and tested with the remaining 90%. It is apparent that dictionary based classification outperformed the SPC and k-means clustering. With the sparse signal representation, the dictionary based clustering was robust to noise as noted in the results, even with a noise level of 0.25, the accuracy was 0.64 ± 0.11. FDDL Dictionary learning used the training samples to learn the space where the given signal could be coded. That is, instead of pre-assuming a particular form of features (e.g. wavelets in SPC), we learned the features from data. Significantly, the learned features are for the specific problem in hand. This is in contrast to PCA, which learns features to optimize a different objective function. This makes dictionary learning problem specific and improves the clustering performance. Another important factor with clustering using FDDL is that both the coding coefficients and reconstruction error are discriminative. Thus, making use of both of them, as we did in (TODOREF 7) and (TODOREF 8), gives more accurate clustering results.

Although overlapping of spike waveforms was not studied in this paper, FDDL dictionary learning has the capability of detecting the overlapped signals and decomposing them as a sparse superposition of several dictionary elements. These dictionary atoms elements are the features for clustering. Building an over complete dictionary [TODOREF 13] provides the space that when used to decode the input signal we can get the mixing weights and the firing time from the sparse coded

coefficients.

IV. CONCLUSION AND FUTURE WORK

A dictionary learning based spike sorting algorithm has been developed, motivated by the powerful reconstruction and discrimination capability of dictionary learning. The FDDL method used the Fisher discrimination criterion to build structured dictionary, whose sub-dictionaries had class specific labels, from the training samples. These sub-dictionaries provided basis for clustering of testing samples. The discriminative ability of FFDL is two fold. First, each learned subdictionary can well reconstruct signals from the corresponding class, and it has poor representation of other classes signals. Second, the sparse coding coefficients are also discriminative. Therefore, the developed clustering criterion used both the discriminative reconstruction error and sparse coding coefficients to cluster the testing spike signals. When tested with simulated data sets and compared to other clustering methods, the method gave high performance in terms of clustering accuracy and reconstruction mean squared error.

As a future direction, the method can be further extended to resolve spike overlapping. The dictionary coding coefficients have the ability to represent the overlapped spikes as a sparse superposition of several dictionary atoms. Thus, we can estimate the time onsets and amplitudes of overlapped spikes. In the ongoing work, we are building an over complete dictionary to test the overlapped signals and apply the FDDL on real data signals.