Signal Processing 153 (2018) 300–310

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Consistent adaptive sequential dictionary learning☆

Abd-Krim Seghouane*, Asif Iqbal

Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, Australia

ARTICLE INFO

Article history:
Received 2 March 2018
Revised 17 June 2018
Accepted 23 July 2018
Available online 29 July 2018

Keywords:
Dictionary learning
Sequential learning
Sparsity
Penalized rank one matrix approximation

ABSTRACT

Algorithms for learning overcomplete dictionaries for sparse signal representation are mostly iterative minimization methods that alternate between a sparse coding stage and a dictionary update stage. For most however, the notion of consistency of the learned quantities has not been addressed. Based on the observation that the observed signals can be approximated as a sum of rank one matrices, a new adaptive dictionary learning algorithm is proposed in this paper. It is derived via sequential adaptive penalized rank one matrix approximation where the ℓ_1-norm is introduced as a penalty promoting sparsity. The proposed algorithm uses a block coordinate descent approach to consistently estimate the unknowns and has the advantage of having simple closed form solutions for both the sparse coding and dictionary update stages. The consistency properties of both the estimated sparse code and dictionary atom are provided. The performance of the proposed algorithm compared to some state of the art algorithms is illustrated on both simulated data and a real functional magnetic resonance imaging (fMRI) data set from a finger tapping experiment.

1. Introduction

Whether it is for restoration, filtering, compression, or higher level tasks such as extracting meaningful features and uncovering structural relationships, transforming signals and images to other domains and the characterizations with some desirable properties constitutes one of the main approach or procedure in signal and image processing. The last few years have seen sparse signal representations with overcomplete dictionary learning methods take an important place in signal and image processing offering solutions that outperform classical approaches in most cases. These methods via sparsity constraints have emerged as fundamental alternatives to the traditional Tikhonov regularization scheme and have proven successful in solving a variety of signal and image processing problems such as image denoising [1], impulse noise removal [2], radar target high resolution range profile (HRRP) recognition [3], acoustic source localization [4], face recognition [5], discrimination [6], seismic data analysis [7], hyperspectral image classification [8], speech separation [9], image super-resolution [10], and functional magnetic resonance imaging (fMRI) data analysis [11,12]. While the underlying key constraint of all dictionary learning methods or algorithms is that the observed signal is sparse within an overcomplete basis, which explicitly means that the signal can be adequately represented using a small number of the available dictionary atoms; their particularity is that the dictionary is also learned or adapted to find the sparse approximation that best describes the observed signals. Given a set of centered signals $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\} \in \mathbb{R}^{n\times N}$ dictionary learning methods learn a dictionary $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_K\} \in \mathbb{R}^{n\times K}$, $N > K > n$, from $\mathbf{Y}$ such that each signal $\mathbf{y} \in \mathbf{Y}$ can be well approximated by a sparse linear combination of $\{\mathbf{d}_i\}_{i=1}^K$; i.e.; $\mathbf{y} = \sum_{i=1}^K x_i \mathbf{d}_i$; where most of the coefficients x_i's are zero or close to zero. This problem is often formulated as follows

$$\min_{\mathbf{D},\mathbf{X}} \| \mathbf{Y} - \mathbf{DX} \|_F^2 \ \text{ s.t. } \ \| \mathbf{x}_i \|_0 \le s, \ \forall\ 1 \le i \le N, \tag{1}$$

where the $\mathbf{x}_i$'s are the column vectors of $\mathbf{X}$, $\|.\|_F$ is the Frobenius norm, $\|.\|_0$ is the l_0 quasi-norm, which counts the number of nonzero coefficients and s denotes the maximum sparsity level allowed for each signal from $\mathbf{Y}$. In addition, the dictionary is usually constrained to belong to the set of feasible dictionaries defined by $D = \{\mathbf{D} \in \mathbb{R}^{n\times K} : \ \| \mathbf{d}_k \|_2 = 1 \ \ \forall k\}$, where $\|.\|_2$ is the l_2 norm and $\mathbf{d}_k$ is the kth column of $\mathbf{D}$ to avoid scaling ambiguity.

Dictionary learning algorithms approximately solve the non-convex and NP-hard problem (1). For most, they consist of two stages: a sparse coding stage and a dictionary update stage. In the first stage the dictionary is kept constant and the sparsity constraint is used to update the codes $\{x_i\}_{i=1}^K$ to produce a sparse linear approximation for the observed signals with the current dictionary. A number of sparse coding algorithms have been

☆ This work was supported by the Australian Research Council through Grant FT. 130101394.

* Corresponding author.
E-mail addresses: Abd-krim.seghouane@unimelb.edu.au (A.-K. Seghouane), Aiqbal1@student.unimelb.edu.au (A. Iqbal).

https://doi.org/10.1016/j.sigpro.2018.07.018
0165-1684/

proposed to perform this stage among them greedy and relaxation algorithms [13]. In the second stage, based on the current sparse codes, the dictionary is updated to minimize a residual cost function generally based on the Frobenius norm subject to a column normalization constraint. This is also a non-convex problem due to the presence of a constraint on the atoms norm. Dictionary learning methods iteratively perform the two stages of sparse coding to find $\mathbf{X}$ and dictionary update to find $\mathbf{D}$ until convergence. The performance difference of existing overcomplete dictionary learning methods is mainly due to differences in the dictionary update stage since most of these methods share a similar sparse coding stage. Besides the difference in the approach used to update the dictionary, the dictionary update can be made sequential where each dictionary atom $\mathbf{d}_i$, $i = 1, \ldots, K$, is updated separately as in [14] for example or in parallel where the dictionary $\mathbf{D}$ is updated at once as in [15]. While the parallel update approach may be preferred for its computational cost advantage, the sequential approach generally offers better performance because it produces finer-tuned dictionaries. In other algorithms, its a constraint to promote incoherence of the learned dictionary $\mathbf{D}$, that has been incorporated in the dictionary update stage [16,17]. Another approach used for obtaining dictionary learning algorithms is to consider the convex relaxation variant of (1) obtained by replacing the ℓ_0-norm for sparsity with the ℓ_1-norm. A number of dictionary learning algorithms based on an ℓ_1-norm sparse coding stage have been proposed, among them online versions such as [18] and batch type version such as [19–22]. The algorithms proposed in [21,22] have the advantage of performing better than some state of the art methods with a substantially lower computational cost. They have the particularity of having a dictionary update stage where the whole row of the sparse coding matrix and associated dictionary atom are updated at the same time instead of updating only its non zero entries as in [14]. Similar to [21–23], these algorithms are based on a penalized matrix decomposition approach [24,25]. Within this framework, dictionary learning algorithms that enforce other type of constraints obtained from a priori knowledge on the signal matrix can be obtained. This is for example the case of [11] where dictionary learning algorithms enforcing smoothness on the dictionary atoms via penalization and basis expansion were proposed for fMRI data analysis. This constraint is used to reflect the a priori knowledge that the fMRI signal at a fixed voxel over time is believed to be smooth and of low frequency. The ℓ_1-norm penalty used in the convex relaxation variant of (1) leads to the classical LASSO method [26] for the sparse coding stage. However, it is well known now that the traditional LASSO may not be fully efficient [27] and its model selection results could be inconsistent [28]. The main reason for this inconvenience is that the LASSO uses the same amount of shrinkage for each coefficient or code x_i, $i = 1, \ldots, K$ when considering the regression of $\mathbf{y}_i$ over the dictionary $\mathbf{D}$. Motivated by this shortcoming of the ℓ_1, a new dictionary learning algorithm which is not only computationally efficient but also has the advantage over the widely used existing algorithms for generating better results is proposed in this paper. It is derived via adaptive sequential penalized rank one matrix approximation where a block coordinate descent approach is used to consistently estimate the vector pairs of the different rank one approximation matrices. The approach adopted for obtaining the proposed algorithm exploits the observation that the observed signals can be approximated as a sum of rank one matrix approximations [21–23]. The adaptive aspect of the algorithm allows different amounts of shrinkage to be used for different entries of the sparse code matrix $\mathbf{X}$ and unlike previous methods, detailed consistency results demonstrating the consistency of the algorithm estimates are provided. Furthermore, the dictionary learning algorithm has the advantage of having simple sparse coding and dictionary update stages as the former corresponds to a vector thresholding step whereas the later corresponds to matrix vector multiplication and each step of the algorithm generates consistent estimates.

We begin by reviewing some background work on dictionary learning in Section 2. The proposed dictionary learning algorithm is described in Section 3 with its consistency properties. Numerical experiments comparing the proposed algorithm with some stat of the art algorithms are presented in Section 4. Concluding remarks are given in Section 5.

2. Background

In this section, we briefly review some background work on ℓ_1-norm based dictionary learning and give some justifications that pave the way for the proposed method.

Given a collection of centered signals $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\} \in \mathbb{R}^{n \times N}$, dictionary learning methods can be obtained by finding a dictionary $\mathbf{D}$ and a sparse representation matrix $\mathbf{X}$ such that

$$\min_{\mathbf{D} \in D, \mathbf{X}} \frac{1}{2} \parallel \mathbf{Y} - \mathbf{D}\mathbf{X} \parallel_F^2 + \lambda \parallel \mathbf{X} \parallel_1, \tag{2}$$

where the $\parallel \mathbf{X} \parallel_1 = \sum_{i=1}^{K} \sum_{j=1}^{N} |x_{ij}|$, $\|.\|_F$ is the Frobenius norm and λ is the regularization or tuning parameter controlling the amount of sparsity. In addition, the dictionary is constrained to belong to the set of feasible dictionaries defined by $D = \{\mathbf{D} \in \mathbb{R}^{n \times K} : \ \parallel \mathbf{d}_k \parallel_2 = 1 \quad \forall k\}$, where $\|.\|_2$ is the l_2 norm and $\mathbf{d}_k$ is the kth column of $\mathbf{D}$ to avoid scaling ambiguity.

A natural approach to solving problem (2) that is adopted by most dictionary learning algorithms is to alternate the minimization between the two variables; i.e.; minimizing over one variable while keeping the other one fixed. This leads to a two stage alternating iteration optimization scheme composed of a sparse coding stage to optimize $\mathbf{X}$ given a fixed dictionary $\mathbf{D}$

$$\min_{\mathbf{X}} \sum_{i=1}^{N} \parallel \mathbf{y}_i - \mathbf{D}\mathbf{x}_i \parallel_F^2 + \lambda \parallel \mathbf{x}_i \parallel_1, \ \forall \ 1 \le i \le N. \tag{3}$$

where $\mathbf{y}_i$ and $\mathbf{x}_i$ are the column vectors of $\mathbf{Y}$ and $\mathbf{X}$, respectively. The sparse coding stage problem (3) consists of a set of ℓ_1-regularized least squares problems. A number of efficient methods are available for solving this type of problem, among them the homotopy method [29] or coordinate descent with soft thresholding [30,31]. Next, the dictionary update stage where $\mathbf{X}$ is fixed and $\mathbf{D}$ is obtained by solving

$$\mathbf{D} = \arg\min_{D} \parallel \mathbf{Y} - \mathbf{D}\mathbf{X} \parallel_F^2 \tag{4}$$

followed by normalizing its columns constitutes the second stage.

Methods used for the dictionary update stage either update the atoms sequentially [14] by breaking the global minimization (4) into K sequential minimization problems or simultaneously all at once [15]. Some additional constraints on $\mathbf{D}$ are sometimes used in applications to improve the performance. One such constraint is an upper bound on the mutual coherence of the dictionary which characterizes the correlation of the dictionary atoms [16,17]. In this widely used two stages approach the objective (1) decreases (or is left unchanged) at each iteration, so that convergence to a local minimum is guaranteed. This makes this procedure strongly dependent on the initial dictionary. The widely used starting method is to use a random collection of K signals from $\mathbf{Y}$ as initial dictionary.

Different methods have been adopted for sequential dictionary learning which is the focus of this paper. Among them, the singular value decomposition (SVD) [14] on a reduced error matrix to generate the atom update as follows

$$\{\mathbf{d}_k, \tilde{\mathbf{x}}_k\} = \arg\min_{\mathbf{d}_k, \tilde{\mathbf{x}}_i^{row}} \parallel \mathbf{Y} - \mathbf{D}\mathbf{X} \parallel_F^2 = \arg\min_{\mathbf{d}_k, \tilde{\mathbf{x}}_k^{row}} \parallel \mathbf{E}_k^R - \mathbf{d}_k \tilde{\mathbf{x}}_k^{row} \parallel_F^2. \tag{5}$$