Signal Processing 123 (2016) 42–52

Contents lists available at ScienceDirect

Signal Processing

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

ℓ_1-K-SVD: A robust dictionary learning algorithm with simultaneous update

Subhadip Mukherjee, Rupam Basu, Chandra Sekhar Seelamantula *

Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

A R T I C L E I N F O

Article history:
Received 1 April 2015
Received in revised form
6 November 2015
Accepted 6 December 2015
Available online 29 December 2015

Keywords:
Dictionary learning
Sparsity
Noise robustness
ℓ_1-minimization
Iteratively reweighted least-squares (IRLS) algorithm
ℓ_1-K-SVD

A B S T R A C T

We develop a new dictionary learning algorithm called the ℓ_1-K-SVD, by minimizing the ℓ_1 distortion on the data term. The proposed formulation corresponds to maximum a posteriori estimation assuming a Laplacian prior on the coefficient matrix and additive noise, and is, in general, robust to non-Gaussian noise. The ℓ_1 distortion is minimized by employing the *iteratively reweighted least-squares* algorithm. The dictionary atoms and the corresponding sparse coefficients are simultaneously estimated in the dictionary update step. Experimental results show that ℓ_1-K-SVD results in noise-robustness, faster convergence, and higher atom recovery rate than the method of optimal directions, K-SVD, and the robust dictionary learning algorithm (RDL), in Gaussian as well as non-Gaussian noise. For a fixed value of sparsity, number of dictionary atoms, and data dimension, ℓ_1-K-SVD outperforms K-SVD and RDL on small training sets. We also consider the generalized ℓ_p, $0<p<1$, data metric to tackle heavy-tailed/impulsive noise. In an image denoising application, ℓ_1-K-SVD was found to result in higher peak signal-to-noise ratio (PSNR) over K-SVD for Laplacian noise. The structural similarity index increases by 0.1 for low input PSNR, which is significant and demonstrates the efficacy of the proposed method.

1. Introduction

Data-independent dictionaries such as discrete cosine transform (DCT), wavelet, curvelet, and contourlet have been extensively used for the applications of image denoising, compression, super-resolution, etc. Fixed dictionaries are generic, offer easy and efficient implementation, but are limited in their ability to provide a parsimonious representation of signals/images. The vast volume of research in compressive sensing [1,2], where one seeks to recover a sparse signal from its noisy linear measurements, has motivated researchers to intensify their quest for suitable dictionaries in which the signal of interest admits a sparse representation. When the dictionaries are *learnt* in a data-dependent or application specific manner, they lead to higher performance than their data-independent counterparts.

In this paper, we propose a dictionary learning approach that is robust to non-Gaussian noise contamination in the dataset, by applying a robust data fidelity metric, namely the ℓ_1-norm. We demonstrate through experiments that the proposed approach not only results in robustness to additive noise, but is also more efficient than the competing algorithms when the size of the training dataset is small. Before proceeding with the developments, we provide a brief overview of some recent literature on dictionary learning.

1.1. Review of relevant literature

A survey of the state-of-the-art dictionary learning techniques is given in [3,4]. Seminal contributions to the problem of data-adaptive dictionary learning were made by Aharon et al. [5,6], who proposed the K-SVD algorithm,

* Corresponding author.
E-mail addresses: subhadip@ee.iisc.ernet.in (S. Mukherjee), rbasu28@ece.iisc.ernet.in (R. Basu), chandra.sekhar@ee.iisc.ernet.in (C.S. Seelamantula).

http://dx.doi.org/10.1016/j.sigpro.2015.12.008
0165-1684/

which, by far, is the most popular algorithm for data-dependent dictionary learning. Theoretical guarantees on the performance of K-SVD are available in [7,8]. Dai et al. [9] developed a generalized framework for dictionary learning, known as *simultaneous codeword optimization* (SimCO). The *method of optimal directions* (MOD) [10] and the K-SVD algorithms emerge as special cases of the generalized formulation.

K-SVD uses the ℓ_2 distortion as a measure of data fidelity and finds applications in image denoising [5], where the noise is zero-mean, additive, and follows a white Gaussian distribution. The training is carried out either on the corpus of clean image patches or on patches derived from the noisy image itself. Since the K-SVD algorithm is not computationally efficient at handling large-dimensional image patches, a sparsity promoting prior is added locally on smaller patches, together with a global prior that enforces proximity between the noisy and the estimated denoised images.

Besides the synthesis model, the dictionary learning problem has also been addressed from the analysis and transform perspectives [11–14]. Apart from denoising, dictionary-based techniques also find applications in image super-resolution, inpainting, image restoration, etc. [15–17]. Discriminative sparse coding (DSC) algorithms such as discriminative K-SVD have been proposed in the literature for solving the signal classification problem [18,19]. Liu et al. [20] proposed the multiview Hessian discriminative sparse coding (MHDSC) algorithm, wherein one combines Hessian regularization with DSC for multiview classification problem. The MHDSC algorithm has been applied successfully for image annotation. Sun et al. [21] recently proposed an algorithm for learning discriminative dictionaries to solve the problem of image classification [22], where a combination of terms measuring discriminative fidelity, sub-dictionary incoherence, and group sparsity is minimized. Algorithms based on dictionary learning are not limited to only image processing applications. They have been widely used in other domains, including single channel speech separation [23], removal of ballistocardiogram effect from the EEG signals [24], human action recognition [25], to name a few. Dictionary learning is also equivalent to *soft clustering* and can be used to effectively learn inherently low-dimensional subspaces from large-dimensional data [26]. From a pattern classification perspective, dictionary learning can be interpreted as a technique to learn a union of low-dimensional subspaces from given training vectors [27].

In many applications, the assumption of Gaussianness of the additive noise may not be accurate, thus rendering ℓ_2-minimization-based algorithms suboptimal. Moreover, ℓ_2 minimization algorithms may lead to oversmoothing, causing loss of image details. One way to address the problem is to develop robust algorithms by minimizing the ℓ_1 data metric. Algorithms for minimizing ℓ_1-based data terms have been proposed in the context of non-negative matrix factorization (NMF) for document detection [28]. Liu and Tao have recently developed an NMF technique with the Manhattan or ℓ_1 distortion metric [29], and carried out systematic analysis of the reconstruction error by decomposing it into approximation error and estimation error, with the latter bounded by the generalization error bound. They also determined rigorously how dimensionality reduction affects both errors. Recently, the problem of robust dictionary learning (RDL) has been addressed by Lu et al. [30] in an online setting, where the training examples are revealed sequentially [31]. Unlike K-SVD, the RDL algorithm does not offer the flexibility to simultaneously update the dictionary and the coefficients over the currently estimated support. A dictionary learning algorithm for impulse noise removal from images, through $\ell_1-\ell_1$ minimization, was proposed by Wang et al. [32], where an augmented Lagrangian framework is used to solve the resulting optimization problem. A notable contribution towards solving the regression problem in non-Gaussian noise was made by Liu and Tao [33], who addressed the regression problem in Cauchy distributed noise. They considered least-squares, least absolute deviation, and Cauchy loss, and showed that they are in the increasing order of robustness to outliers.

1.2. This paper

We develop a dictionary learning algorithm aimed at minimizing the ℓ_1 distortion on the data term. The ℓ_1 error is minimized using iteratively reweighted least-squares (IRLS) technique [34–36], employing which we solve the ℓ_1 minimization problem by iteratively solving a series of ℓ_2 minimization problems (cf. Section 2). The IRLS approach is appealing, because it enables one to minimize the ℓ_1 cost in an iterative manner, where in each iteration, one essentially solves a weighted ℓ_2 minimization problem, which has a closed-form solution. Consequently, the proposed method possesses the simplicity of the ℓ_2 minimization problem, and the noise-robustness of the ℓ_1 data penalty. The idea of reweighting is effective for developing computationally efficient algorithms for otherwise hard optimization problems. The notion of reweighting is also gaining importance in machine learning and pattern classification [37]. The use of ℓ_1 metric for data fidelity results in robustness for suppression of heavy-tailed noise from images while preserving structure, measured in terms of structural similarity index (SSIM), as demonstrated by the experimental results (cf. Section 3). Unlike the RDL algorithm [30], in ℓ_1-K-SVD, the dictionary atoms and the entries of the coefficient matrix are simultaneously updated (similar to K-SVD), offering better convergence performance. We show that the formulation also generalizes to the case of the ℓ_p metric, where $0<p<1$, which is known to promote sparsity better than the ℓ_1 norm. The resulting ℓ_p-K-SVD algorithm outperforms the K-SVD when the training data is corrupted by impulsive noise. Thus, the proposed method provides a unified framework to incorporate different penalties on the data term, customized based on the distribution of the additive noise.

2. Proposed dictionary learning approach

Denote the training dataset by $\mathcal{T}$, which contains N exemplars $\{\mathbf{y}_n\}_{n=1}^N$ in $\mathbb{R}^m$, corrupted by additive noise. The objective is to learn a dictionary D containing K atoms, tailored to $\mathcal{T}$, such that D represents the signals in $\mathcal{T}$ using sparse coefficient vectors $\mathbf{x}_n$. Typically, D is overcomplete, that is, $m<K$. We use the symbols Y and X to denote the