Artificial Intelligence 238 (2016) 83–95

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

# Smooth sparse coding via marginal regression for learning sparse representations ☆

Krishnakumar Balasubramanian [a],*, Kai Yu [b], Guy Lebanon [c]

[a] *Department of Statistics, University of Wisconsin-Madison, USA*
[b] *Horizong Robotics, Beijing, China*
[c] *LinkedIn, USA*

ARTICLE INFO

*Article history:*
Received 16 December 2013
Received in revised form 9 April 2016
Accepted 14 April 2016
Available online 16 May 2016

*Keywords:*
Sparse coding
Dictionary learning
Vision

ABSTRACT

We propose and analyze a novel framework for learning sparse representations based on two statistical techniques: kernel smoothing and marginal regression. The proposed approach provides a flexible framework for incorporating feature similarity or temporal information present in data sets via non-parametric kernel smoothing. We provide generalization bounds for dictionary learning using smooth sparse coding and show how the sample complexity depends on the $L_1$ norm of kernel function used. Furthermore, we propose using marginal regression for obtaining sparse codes which significantly improves the speed and allows one to scale to large dictionary sizes easily. We demonstrate the advantages of the proposed approach, both in terms of accuracy and speed by extensive experimentation on several real data sets. In addition, we demonstrate how the proposed approach can be used for improving semi-supervised sparse coding.

## 1. Introduction

Sparse coding is a popular unsupervised paradigm for learning sparse representations of data samples that are subsequently used in classification tasks. In standard sparse coding, each data sample is coded independently with respect to the dictionary. We propose a smooth alternative to traditional sparse coding that incorporates feature similarity, temporal similarity or similar user-specified similarity measures between the samples into the coding process.

The idea of smooth sparse coding is motivated by the relevance weighted likelihood principle. Our approach constructs a code that is efficient in a smooth sense and as a result leads to improved statistical accuracy over traditional sparse coding. The smoothing operation, which can be expressed as non-parametric kernel smoothing, provides a flexible framework for incorporating several types of domain information that might be available for the user. For example, in image classification, one could use: (1) kernels in feature space for encoding similarity information for images and videos and (2) kernels in time space in case of videos for incorporating temporal relationship. Apart from this, the kernel could also be used to encode similarity information in semi-supervised learning setting.

Most sparse coding training algorithms fall under the general category of alternating procedures with a convex lasso regression sub-problem. While efficient algorithms for such cases exist [17], their scalability for large dictionaries remains a challenge. We propose a novel training method for sparse coding based on marginal regression, rather than solving the

☆ This paper is an invited revision of a paper which first appeared at the International Conference on Machine Learning (ICML) 2013.

* Corresponding author.
*E-mail addresses:* krizna@stat.wisc.edu (K. Balasubramanian), yukai@baidu.com (K. Yu), glebanon@gmail.com (G. Lebanon).

http://dx.doi.org/10.1016/j.artint.2016.04.009
0004-3702/© 2016 Elsevier B.V. All rights reserved.

traditional alternating method with lasso sub-problem. Marginal regression corresponds to performing univariate linear regression corresponding to each dimension, followed by a thresholding step to promote sparsity. For large dictionary sizes, this leads to a significant speedup compared to traditional sparse coding methods without sacrificing statistical accuracy.

Note that the notion of speedup we mention should be interpreted appropriately. There are two contributions we make: (i) the smoothing operation and (ii) the use of marginal regression updates in places of lasso updates. It should be noted that the smoothing operation requires additional computation. The source of speedup is replacing the lasso step with marginal regression step in the alternating minimization procedure. For a fair comparison, one needs to look at the performance of smooth sparse coding (or standard sparse coding) when it uses lasso updates and marginal regression updates. We report the overall timing comparison between the different methods in the experimental section for clarification.

We also develop theoretical analysis that extends the sample complexity result of [29] for dictionary learning using standard sparse coding to the smooth sparse coding case. This result specifically shows how the sample complexity depends on the $L_1$ norm of the kernel function used. Our contributions lead to improved classification accuracy in conjunction with significant computational speedup. Below we summarize our main contributions:

1. we propose a framework based on kernel-smoothing for incorporating feature, time or other similarity information between the samples into sparse coding.
2. we provide sample complexity results for dictionary learning using smooth sparse coding.
3. we propose an efficient marginal regression training procedure for sparse coding.
4. We successfully apply the proposed method in various classification tasks and report improved performance in several situations.

## 2. Related work

Our approach is related to the local regression method [19,11]. More recent related work is [21] that uses smoothing techniques in high-dimensional lasso regression in the context of temporal data. Another recent approach [34], achieves code locality by approximating data points using a linear combination of nearby basis points. The main difference is that traditional local regression techniques [19,11,21] do not involve basis learning. In this work, we propose to learn the basis or dictionary along with the regression coefficients locally. This could be viewed as a high dimensional generalization of low dimensional local smoothing problems with no basis learning. Here we argue that one could directly learn the basis simultaneously while using traditional local-smoothing techniques for improved performance. Furthermore, it provides a natural way to incorporate various similarity information constructed from the data samples themselves in to the sparse coding process.

In contrast to previous sparse coding papers we propose to use marginal regression for learning the regression coefficients, which results in a significant computational speedup with no loss of accuracy. Marginal regression is a relatively old technique that has recently reemerged as a computationally faster alternative to lasso regression [8]. See also [10] for a statistical comparison of lasso regression and marginal regression.

## 3. Smooth sparse coding

**Notation:** We use lower case letters, for example $x$, to represent vectors and upper case letters, for example $X$, to represent matrices, in appropriately defined dimensions. We use $\|\cdot\|_p$ to represent the $L_p$ norm of a vector (we use mostly use $p = 1, 2$ in this paper), $\|\cdot\|_F$ to represent the Frobenius norm of a matrix and $|f|_p$ to represent $L_p$ norm of the function $f$ defined as $(\int |f|^p \, d\mu)^{1/p}$. Data samples are denoted by subscripts, for example $\{x_i\}_{i=1}^n$ corresponds to $n$ data samples, where each sample $x_i$ is a $d$-dimensional vector.

The standard sparse coding problem consists of solving the following optimization problem,

$$\begin{aligned}
&\min_{\substack{D\in\mathbb{R}^{d\times K}\\ \beta_i\in\mathbb{R}^K, i=1,\ldots,n}} && \sum_{i=1}^{n} \|x_i - D\beta_i\|_2^2 \\
&\text{subject to} && \|d_j\|_2 \le 1 \quad j = 1,\ldots K \\
& && \|\beta_i\|_1 \le \lambda \quad i = 1,\ldots n,
\end{aligned}$$

where $\beta_i \in \mathbb{R}^K$ corresponds to the encoding of sample $x_i$ with respected to the dictionary $D \in \mathbb{R}^{d\times K}$ and $d_j \in \mathbb{R}^d$ denotes the $j$-column of the dictionary matrix $D$. The dictionary is typically over-complete, implying that $K > d$.

Object recognition is a common sparse coding application where $x_i$ corresponds to a set of features obtained from a collection of image patches, for example Scale-Invariant Feature Transform (SIFT) features [20]. The dictionary $D$ corresponds to an alternative coding scheme that is higher dimensional than the original feature representation. The $L_1$ constraint promotes sparsity of the new encoding with respect to $D$. Thus, every sample is now encoded as a sparse vector that is of higher dimensionality than the original representation.

In some cases the data exhibits a structure that is not captured by the sparse coding setting. For example, SIFT features corresponding to samples from the same class are presumably closer to each other compared to SIFT features from other