Signal Processing 130 (2017) 93–104

Contents lists available at ScienceDirect

Signal Processing

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

Sparse Bayesian dictionary learning with a Gaussian hierarchical model ☆

Linxiao Yang [a], Jun Fang [a,*], Hong Cheng [b], Hongbin Li [c]

[a] National Key Laboratory on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] School of Automation, University of Electronic Science and Technology of China, Chengdu 611731, China
[c] Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA

ARTICLE INFO

Article history:
Received 8 July 2015
Received in revised form
14 June 2016
Accepted 16 June 2016
Available online 21 June 2016

Keywords:
Dictionary learning
Gaussian-inverse Gamma prior
Variational Bayesian
Gibbs sampling

ABSTRACT

We consider a dictionary learning problem aimed at designing a dictionary such that the signals admit a sparse or an approximate sparse representation over the learnt dictionary. The problem finds a variety of applications including image denoising, feature extraction, etc. In this paper, we propose a new hierarchical Bayesian model for dictionary learning, in which a Gaussian-inverse Gamma hierarchical prior is used to promote the sparsity of the representation. Suitable non-informative priors are also placed on the dictionary and the noise variance such that they can be reliably estimated from the data. Based on the hierarchical model, a variational Bayesian method and a Gibbs sampling method are developed for Bayesian inference. The proposed methods have the advantage that they do not require the knowledge of the noise variance a priori. Numerical results show that the proposed methods are able to learn the dictionary with an accuracy better than existing methods, particularly for the case where there is a limited number of training signals.

1. Introduction

Sparse representation has been of significant interest over the past few years. It has found a variety of applications in practice as many natural signals admit a sparse or approximately sparse representation in a certain basis [1–3]. In many applications such as image denoising and interpolation, signals often have a sparse representation over a pre-specified non-adaptive dictionary, e.g. discrete cosine/wavelet transform (DCT/DWT) bases. Nevertheless, recent research [4,5] has shown that the recovery, denoising and classification performance can be considerably improved by utilizing an adaptive dictionary that is learnt from training signals [5,6]. This has inspired studies on dictionary learning aimed to design overcompelete dictionaries that can better represent the signals. A number of algorithms, such as the K-singular value decomposition (K-SVD) [4], the method of optimal directions (MOD) [7], dictionary learning with the majorization method [8], and the simultaneous codeword optimization (SimCO) [9], were developed for overcomplete dictionary learning and sparse representation. Most algorithms formulate the dictionary learning as an optimization problem which is solved via a two-stage iterative process, namely, a sparse coding stage and a dictionary update stage. The main difference among these algorithms lies in the dictionary update stage. Specifically, the MOD method [7] updates the dictionary via solving a least square problem which admits a closed-form solution for dictionary update. The K-SVD algorithm [4], instead, updates the atoms of the dictionary in a sequential manner and while updating each atom, the atom is updated along with the nonzero entries in the corresponding row vector of the sparse matrix. The idea of sequential atom update was later extended to provide sequential update of multiple atoms each time [9], and recently generalized to parallel atom-updating in order to further accelerate the convergence of the iterative process [10]. These methods [4,7–10], although offering state-of-the-art performance, have several limitations. Specifically, they may require the knowledge of the sparsity level or the noise/residual variance for sparse coding (e.g. [4]), or this knowledge is needed for meticulously selecting some regularization parameters to properly control the tradeoff between the sparsity level and the data fitting error (e.g. [8,10]). In practice, however, the prior information about the noise variance and sparsity level is usually unavailable and an inaccurate estimation may result in substantial performance degradation. To mitigate these limitations, a nonparametric Bayesian dictionary learning method called beta-Bernoulli process factor analysis (BPFA) was recently developed in [11]. The proposed method can estimate the usage frequency of each atom, based on which the required number of atoms can be automatically inferred. Moreover, BPFA is also able to

☆This work was supported in part by the National Science Foundation of China under Grants 61428103, 61522104, and the National Science Foundation under Grant ECCS-1408182.

* Corresponding author.
E-mail addresses: JunFang@uestc.edu.cn, 201321190224@std.uestc.edu.cn (J. Fang), hcheng@uestc.edu.cn (H. Cheng), Hongbin.Li@stevens.edu (H. Li).

http://dx.doi.org/10.1016/j.sigpro.2016.06.016
0165-1684/

automatically infer the noise variance from the test image. These merits are deemed an important advantage over other dictionary learning methods. For [11], the posterior distributions cannot be derived analytically, and a Gibbs sampler was used for Bayesian inference. We also note that a class of online dictionary learning algorithms were developed in [12–16]. Unlike the above batch-based algorithms [4,7,9,10] which use the whole set of training data for dictionary learning, online algorithms continuously update the dictionary using only one or a few (or a small amount of) training data, which enables them to handle very large data sets.

In this paper, we propose a new hierarchical Bayesian model for dictionary learning, in which a Gaussian-inverse Gamma hierarchical prior [17,18] is used to promote the sparsity of the representation. Suitable non-informative priors are also placed on the dictionary and the noise variance such that they can be reliably inferred from the data. Based on the hierarchical model, a variational Bayesian method [19–21] and a Gibbs sampling method [22] are developed for Bayesian inference. For both inference methods, there are two different ways to update the dictionary: we can either update the whole set of atoms in one iteration, or update the atoms in a sequential manner. When updating the dictionary as a whole, the proposed variational Bayesian method has a dictionary update formula similar to the MOD method. For the Gibbs sampler, a sequential update seems to be able to expedite the convergence rate and helps achieve additional performance gain. Simulation results show that the proposed Gibbs sampling algorithm has notable advantages over other state-of-the-art dictionary learning methods in a number of interesting scenarios.

Note that the Gaussian-inverse Gamma hierarchical prior used in our paper is quite different from the beta-Bernoulli (also referred to as the spike-and-slab) prior employed in [11]. These two priors have their respective merits and both are widely used to promote the sparsity of solutions. In particular, the use of the Gaussian-inverse Gamma prior for sparse Bayesian learning has achieved great success in the framework of compressed sensing, e.g. [23–26]. It is therefore interesting to examine the problem of dictionary learning with such a prior and see if an additional performance improvement can be achieved. Note that the sparsity-promoting prior model (i.e. the hierarchical Gaussian-inverse Gamma prior) employed in this paper was also used in the sparse PCA framework (e.g. [27]). Nevertheless, to our best knowledge, our paper presents a first attempt to use the hierarchical Gaussian-inverse Gamma prior model to solve the dictionary learning problem. Although dictionary learning is closely related to sparse PCA [27], they still are two different problems with very distinct objectives: dictionary learning tries to learn an overcomplete dictionary to sparsely represent the observed data, whereas the sparse PCA aims to find a few sparse principle components of the underlying data matrix. Also, although sharing some degree of similarity, the prior model used in our paper is not exactly the same as the prior model in [27]. As a consequence, the derivations, update rules, and choice of model parameters in our work are different from those in [27]. Our work also provides an interesting comparison between two different inference methods, namely, the variational Bayes and the Gibbs sampling, for dictionary learning.

The rest of the paper is organized as follows. In Section 2, we introduce a hierarchical prior model for dictionary learning. Based on this hierarchical model, a variational Bayesian method and a Gibbs sampler are developed in Sections 3 and 4 for Bayesian inference. Simulation results are provided in Section 5, followed by concluding remarks in Section 6.

2. Hierarchical model

Suppose we have L training signals $\{\boldsymbol{y}_l\}_{l=1}^L$, where $\boldsymbol{y}_l \in \mathbb{R}^M$. Dictionary learning aims at finding a common sparsifying dictionary $\boldsymbol{D} \in \mathbb{R}^{M\times N}$ such that these L training signals admit a sparse representation over the overcomplete dictionary $\boldsymbol{D}$, i.e.

$$\boldsymbol{y}_l = \boldsymbol{D}\boldsymbol{x}_l + \boldsymbol{w}_l \quad \forall\, l, \tag{1}$$

where $\boldsymbol{x}_l$ and $\boldsymbol{w}_l$ denote the sparse vector and the residual/noise vector, respectively. Define $\boldsymbol{Y} \triangleq [\boldsymbol{y}_1 \ \cdots \ \boldsymbol{y}_L]$, $\boldsymbol{X} \triangleq [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_L]$, and $\boldsymbol{W} \triangleq [\boldsymbol{w}_1 \ \cdots \ \boldsymbol{w}_L]$. The model (1) can be re-expressed as

$$\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X} + \boldsymbol{W}. \tag{2}$$

Also, we write $\boldsymbol{D} \triangleq [\boldsymbol{d}_1 \ \cdots \ \boldsymbol{d}_N]$, where each column of the dictionary, $\boldsymbol{d}_n$, is called an atom.

In the following, we develop a Bayesian framework for learning the overcomplete dictionary and sparse vectors. To promote sparse representations, we assign a two-layer hierarchical Gaussian-inverse Gamma prior to $\boldsymbol{X}$. The Gaussian-inverse Gamma prior is one of the most popular sparsity-promoting priors which has been widely used in compressed sensing [23,24,28]. Specifically, in the first layer, $\boldsymbol{X}$ is assigned a Gaussian prior distribution

$$p(\boldsymbol{X}|\boldsymbol{\alpha}) = \prod_{n=1}^{N}\prod_{l=1}^{L} p(x_{nl}) = \prod_{n=1}^{N}\prod_{l=1}^{L} \mathcal{N}(x_{nl}|0,\, \alpha_{nl}^{-1}), \tag{3}$$

where x_{nl} denotes the $(n,\, l)$th entry of $\boldsymbol{X}$, and $\boldsymbol{\alpha} \triangleq \{\alpha_{nl}\}$ are non-negative sparsity-controlling hyperparameters. The notation $\mathcal{N}(x_{nl}|0,\, \alpha_{nl}^{-1})$ denotes Gaussian distribution with zero mean and variance α_{nl}^{-1}. The second layer specifies Gamma distributions as hyperpriors over the hyperparameters $\{\alpha_{nl}\}$, i.e.

$$p(\boldsymbol{\alpha}) = \prod_{n=1}^{N}\prod_{l=1}^{L} \mathrm{Gamma}(\alpha_{nl};\, a,\, b) = \prod_{n=1}^{N}\prod_{l=1}^{L} \Gamma(a)^{-1} b^a \alpha_{nl}^{a-1} e^{-b\alpha_{nl}}, \tag{4}$$

where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ is the Gamma function. Here the notation $\mathrm{Gamma}(\alpha_{nl};\, a,\, b)$ denotes the Gamma distribution of α_{nl} with parameters a and b. To illustrate the sparsity-promoting property of the Gaussian-inverse Gamma prior, we integrate out the hyperparameter α_{nl} and obtain the marginal distribution of x_{nl}, which was shown to be a student-t distribution, i.e.

$$\begin{aligned} p(x_{nl}) &= \int p(x_{nl}|\alpha_{nl})p(\alpha_n l;\, a,\, b)d\alpha_{nl} \\ &= \frac{b^a\Gamma(a+0.5)}{(2\pi)^{1/2}\Gamma(a)}\left(b + \frac{x_{nl}^2}{2}\right)^{-(a+0.5)}. \end{aligned} \tag{5}$$

When b is very small, say $b = 10^{-6}$, the student-t distribution can be reduced to

$$p(x_{nl}) \propto \left(\frac{1}{x_{nl}^2}\right)^{(a+0.5)}. \tag{6}$$

We can easily see that (6) is a sparsity-promoting prior. Fig. 1 plots the student-t distributions with different choices of a and b. We see that the distribution has a sharp peak around zero when b is sufficiently small. Also, a larger a results in a sharper peak, which implies that a larger a leads to a more sparsity-encouraging prior. In our paper, the parameters a and b are chosen to be $a = 0.5$ and $b = 10^{-6}$.

In addition, in order to prevent the entries in the dictionary from becoming infinitely large, we assume that the atoms of the dictionary $\{\boldsymbol{d}_n\}$ are mutually independent, and upon each atom we place a Gaussian prior, i.e.

$$p(\boldsymbol{D}) = \prod_{n=1}^{N} p(\boldsymbol{d}_n) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{d}_n|\boldsymbol{0},\, \beta\boldsymbol{I}), \tag{7}$$

where β is a parameter whose choice will be discussed later. The noise $\{\boldsymbol{w}_l\}$ are assumed independent multivariate Gaussian noise with zero mean and covariance matrix $(1/\gamma)\boldsymbol{I}$, where the noise