SPARSE MODELING WITH UNIVERSAL PRIORS AND LEARNED INCOHERENT DICTIONARIES

By

Ignacio Ramírez

Federico Lecumberry

and

Guillermo Sapiro

IMA Preprint Series # 2279

(September 2009)

INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA
400 Lind Hall
207 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066 Fax: 612/626-7370
URL: http://www.ima.umn.edu

Sparse Modeling with Universal Priors and Learned Incoherent Dictionaries

Ignacio Ramírez
University of Minnesota
ramir048@umn.edu,

Federico Lecumberry
Universidad de la República
fefo@fing.edu.uy

Guillermo Sapiro
University of Minnesota
guille@umn.edu

September 9, 2009

Abstract

Sparse data models have gained considerable attention in recent years, and their use has led to state-of-the-art results in many signal and image processing tasks. The learning of sparse models has been mostly concerned with adapting the dictionary to tasks such as classification and reconstruction, optimizing extrinsic properties of the trained dictionaries. In this work, we first propose a learning method aimed at enhancing both extrinsic and *intrinsic* properties of the dictionaries, such as the mutual and cumulative coherence and the Gram matrix norm, characteristics known to improve the efficiency and performance of sparse coding algorithms. We then use tools from information theory to propose a sparsity regularization term which has several desirable theoretical and practical advantages over the more standard ℓ_0 or ℓ_1 ones. These new sparse modeling components lead to improved coding performance and accuracy in reconstruction tasks.

1 Introduction

Sparse modeling calls for constructing a succinct representation of some data as a combination of a few typical patterns (atoms) learned from the data itself. Significant contributions to the theory and practice of learning such collections of atoms (usually called dictionaries or codebooks), e.g., [1, 12, 24], and of representing the actual data in terms of them, e.g., [6, 7, 8], have been developed in recent years, leading to state-of-the-art results in many signal and image processing tasks [10, 18, 22, 20, 25]. We refer the reader for example to [3] for a recent review on the subject.

In all cases, the actual dictionary plays a critical role. Current techniques for obtaining such dictionaries involve the optimization of their *extrinsic* properties in terms of the task to be performed (i.e., representation [12], denoising [10, 22], or classification [20]). However, theoretical results addressing the success in recovering sparse signals, as well as the efficiency of sparse coding algorithms, are related to *intrinsic* properties of the dictionary such as the *mutual coherence*, the *cumulative coherence*, and the Gram matrix norm of the dictionary

[9, 30]. We will provide precise definitions of these magnitudes in the sequel. Addressing these important intrinsic properties is one of the goals of the work here presented.

A critical component of sparse modeling is the actual sparsity of the solution, which is controlled by some critical model parameters. Choosing the optimal values of these parameters for the actual signals to model and the problem at hand is a challenging task. Several solutions to this problem have been proposed, ranging from the automatic tuning of the parameters [15] to Bayesian hierarchical models, where these parameters are themselves considered as random variables [14, 15, 31]. In this paper we address this challenge, and at the same time further generalize the standard sparsifying penalty functions, exploiting tools from information theory.

The first contribution of this work is the explicit incorporation of a new term in the sparse modeling formulation that induces the desired intrinsic properties of low mutual coherence and low Gram matrix norm in the learned dictionaries. We show how this leads to the better performance of coding algorithms. The learned dictionaries also exhibit desirable extrinsic properties such as reduced overfitting, a direct consequence of the reduced coherence between the dictionary atoms. Our second contribution is the substitution of the traditional ℓ_0 or ℓ_1 sparsity-inducing priors by one which we derive using information-theoretic tools. This prior has several desirable theoretical and practical properties such as statistical consistency, improved robustness to outliers in the data, and leads to a better sparse reconstruction than ℓ_0 and ℓ_1-based techniques in practice.

The rest of this paper is organized as follows: in Section 2 we introduce the standard framework of sparse modeling. Sections 3 to 5 are dedicated to the derivation of our proposed model. Section 6 gives details on the implementation of the learning algorithm. Sections 7 and 8 present experimental results showing the importance of the proposed sparse model for image representation. Concluding remarks are given in Section 9.

2 Sparse modeling

Let $\mathbf{X} \in \mathbb{R}^{n \times N}$ be a set of N column data vectors $\mathbf{X}_j \in \mathbb{R}^n$, $\mathbf{D} \in \mathbb{R}^{n \times K}$ be a dictionary of K atoms represented as columns $\mathbf{D}_k \in \mathbb{R}^n$, and $\mathbf{A} = \{\alpha_{ij}\} \in \mathbb{R}^{K \times N}, \mathbf{A}_j \in \mathbb{R}^K$, be a set of reconstruction coefficients such that $\mathbf{X} = \mathbf{DA}$. For each $j = 1, \ldots, N$ we define the *active set* of $\mathbf{A}_j$ as $\mathcal{A}_j = \{i : \alpha_{ij} \neq 0\}$, and $\|\mathbf{A}_j\|_0 = |\mathcal{A}_j|$ as its cardinalty. The goal of *sparse modeling* is to design a dictionary $\mathbf{D}$ such that $\mathbf{X} = \mathbf{DA}$ with $\|\mathbf{A}_j\|_0$ sufficiently small (usually below some threshold) for all or most data samples $\mathbf{X}_j$. For a fixed $\mathbf{D}$, the actual computation of $\mathbf{A}$ is called *sparse coding* (SC).

We begin our discussion with the standard ℓ_1 *penalty* modeling problem,

$$(\mathbf{A}^*, \mathbf{D}^*) = \arg\min_{\mathbf{A},\mathbf{D}} \left\{ \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \|\mathbf{A}\|_1 \right\} \tag{1}$$

where $\|\cdot\|_F$ denotes Frobenius norm. The cost function to be minimized in (1) consists of a quadratic *fitting term* and an ℓ_1 *regularization term* for each column of $\mathbf{A}$, the balance of the two being defined by the *penalty parameter* λ. The ℓ_1 norm is used as an approximation to ℓ_0, making the problem convex in $\mathbf{A}$ while still encouraging sparse solutions [3]. Furthermore, it has been shown that under certain conditions on $\mathbf{D}$ and $\mathbf{X}$, the solutions to the ℓ_1-based and ℓ_0-based sparse coding problems coincide, e.g., [4]. The dictionaries learned with the model here proposed explicitly encourage such conditions.

3 Learning incoherent dictionaries

Most techniques for dictionary learning are concerned with the extrinsic properties of the learned dictionaries, e.g., requiring small reconstruction errors while simultaneously imposing a sparsity constraint [10, 12, 22, 24]. However, recent results [3, 7, 30] in sparse coding theory indicate that certain intrinsic properties of the dictionaries have a direct impact on the performance of coding algorithms. Optimizing for such intrinsic properties is the goal of this section.

In the context of compressed sensing, the idea of learning the sensing matrix, not quite the dictionary as here proposed, targeted at the optimization of intrinsic (effective) dictionary properties, has been recently proposed in [9] (see also [28]). It has been also shown that designing the sensing matrix can accelerate the sparse coding optimization [17]. In this section we propose to learn dictionaries that in addition to achieving small reconstruction errors with sparse representations, have intrinsic properties that lead to important performance improvements.

Let us assume that the atoms are normalized, $\|\mathbf{D}_k\| = 1$. The *L-cumulative coherence* of $\mathbf{D}$ is defined as $\bar{\mu}_L(\mathbf{D}) \triangleq \max\left\{\max_{i \notin J} \sum_{j \in J} |\mathbf{D}_i^T \mathbf{D}_j| : J \subseteq \{1, \ldots, K\}, |J| = L\right\}$ [30]. When L is not specified, we assume it to be its maximum value $K-1$, so that $\bar{\mu}(\mathbf{D}) = \bar{\mu}_{K-1}(\mathbf{D})$. Assuming there is a true sparse representation for the data, the success in recovering it by approximating the ℓ_0 sparse formulation using Orthogonal Matching Pursuit (OMP) [23], or by solving the Basis Pursuit (BP) [6] problem [30], is influenced by these quantities, which we optimize for below.

Another important property of $\mathbf{D}$ is its Gram matrix $\mathbf{G} = \mathbf{D}^T\mathbf{D}$, whose matrix ℓ_2-norm $\rho(\mathbf{D})$ is known to influence the speed of convergence of popular shrinkage-based coding algorithms [7, 11]. We recall that $\rho(\mathbf{D}) = \|\mathbf{G}\|_2 = \max\{|\gamma| : \gamma \in \gamma(\mathbf{G})\}$, where $\gamma(\mathbf{G})$ denotes the set of eigenvalues of $\mathbf{G}$. Let $\gamma_{\max}$ denote the eigenvalue with largest absolute value, that is, $\rho(\mathbf{D}) = |\gamma_{\max}|$. By the Gershgorin Circle Theorem [16, pp. 320–321], we have that $\gamma(\mathbf{G}) \subseteq \cup_{k=1}^K \mathcal{C}_k$, where $\mathcal{C}_k = \{y : |y - g_{kk}| \leq \sum_{r \neq k} |g_{kr}|\}$ and $g_{kr} = \mathbf{D}_k^T \mathbf{D}_r$ are the elements of $\mathbf{G}$. For normalized atoms we have for all k that $g_{kk} = 1$, and thus $\sum_{r \neq k} |g_{kr}| \leq \bar{\mu}(\mathbf{D})$. Plugging these values in the Gershgorin Theorem we obtain that $\rho(\mathbf{D}) = |\gamma_{\max}| \leq 1 + \bar{\mu}(\mathbf{D})$. Therefore, we can simultaneously reduce $\rho(\mathbf{D})$ and $\bar{\mu}(\mathbf{D})$ by imposing small off-diagonal elements in $\mathbf{G}$. This suggests the addition of the following term to Equation (1)

$$\left\|\mathbf{D}^T\mathbf{D} - \mathbf{I}_K\right\|_F^2, \tag{2}$$

where $\mathbf{I}_K$ is the $K \times K$ identity matrix. In Section 8 we will show that this new term helps to improve reconstruction accuracy and speed. The final proposed cost function is shown later in Section 5 after a further modification to Equation (1), which we discuss next.

4 Universal models for sparse coding

Sparse models are often presented in a probabilistic framework [19, 24]. From this point of view, the solution for $\mathbf{D}^*$ in (1) can be seen as an approximate solution to a Maximum Likelihood estimation of $\mathbf{D}$ given $\mathbf{X}$, that is, with a slight abuse of notation,

$$\mathbf{D}^* = \arg\max_{\mathbf{D}} \left\{ p(\mathbf{X}|\mathbf{D}) = \int_\Omega p(\mathbf{X}, \mathbf{A}|\mathbf{D}) d\mathbf{A} = \int_\Omega p(\mathbf{X}|\mathbf{A}, \mathbf{D}) p(\mathbf{A}|\mathbf{D}) d\mathbf{A} \right\}, \tag{3}$$

where Ω represents the space where $\mathbf{A}$ takes its values. In this framework, $\mathbf{A}$ is considered to be a *hidden state variable* which is marginalized in the integral. A direct solution of the integral in (3) is generally not possible, so approximate solutions are sought instead. One such approximation, followed in [24], is to maximize the *mode* of the integrand $p(\mathbf{X},\mathbf{A}|\mathbf{D})$ instead of the whole integral, which occurs at $(\mathbf{X},\mathbf{A}^*)$. This is roughly justified by assuming that $p(\mathbf{X},\mathbf{A}|\mathbf{D})$ is unimodal and highly peaked, so that its value at the mode accounts for most of the value of the integral. The maximization is then carried on in the logarithmic scale, resulting in

$$\mathbf{D}^* = \arg\max_{\mathbf{D}} \left\{ \max_{\mathbf{A}} \left\{ \log p(\mathbf{X}|\mathbf{A},\mathbf{D}) + \log p(\mathbf{A}|\mathbf{D}) \right\} \right\}. \tag{4}$$

The formulation (1) is then obtained from (4) considering the reconstruction error to be IID Gaussian with mean 0 and variance σ^2, $p(\mathbf{X}|\mathbf{A},\mathbf{D}) \propto \exp(-\frac{1}{2\sigma^2}\left\|\mathbf{X}-\mathbf{D}\mathbf{A}\right\|_2^2)$; and an IID Laplacian prior with mean 0 and parameter θ on the reconstruction coefficients, which is furthermore assumed to be independent of $\mathbf{D}$, $p(\mathbf{A}|\mathbf{D}) \sim p(\mathbf{A}|\theta) \propto e^{-\theta\|\mathbf{A}\|_1}$. Equation (1) follows by taking the logarithms of both priors and factorizing σ^2 into $\lambda = 2\sigma^2\theta$.

Considering $\mathbf{D}$ fixed we now turn to the problem of finding the aforementioned mode of $p(\mathbf{X},\mathbf{A}|\mathbf{D})$. For a Laplacian prior, this is the sparse coding problem that appears in (1) when optimizing only over $\mathbf{A}$. When considering the statistical modeling of small patches from natural images, the assumption that transform coefficients (e.g., DCT or wavelet) are well modeled by a Laplacian distribution is widely accepted. Assuming the coefficients of $\mathbf{A}$ to be IID leads to the sparse coding problem in (1), as discussed above. Even under these strong assumptions, the selection of an optimal value of λ (or θ for known σ^2), is already a challenging problem (see [15] for example).

It can be argued that a better model would be one with different Laplacian parameters for coefficients associated to different atoms (that is, different rows of $\mathbf{A}$), as it happens with DCT coefficients associated with different basis functions. Even so, an IID assumption for the rows of $\mathbf{A}$ also seems inadequate for natural images, whose statistics vary across different regions (e.g., textures, boundaries). Therefore, instead of tuning for a fixed λ or multiple values of λ, we look for a more general prior that can fit, without knowing $\mathbf{A}$ in advance, almost as well as a Laplacian whose parameter θ was tuned for that specific instance of $\mathbf{A}$. The answer to such problem is given by the information-theoretic concept of *universal coding* which lies at the core of the Minimum Description Length (MDL) principle [2]: given a set of probability models $\mathcal{M} = \{p(\cdot|\theta) : \theta \in \Theta\}$ parameterized by some θ, it is possible to construct a probability model $q(\cdot)$ that fits the data to be modeled approximately as well as the probability model $p(\cdot|\hat{\theta})$ in $\mathcal{M}$ that best fits the data. Here $\hat{\theta}$ is the Maximum Likelihood Estimator (MLE) of θ. One way to construct such a universal prior is through a *Bayesian mixture*. In a Bayesian mixture $q(\cdot)$, the probability of a given coefficient value $q(\alpha) = \Pr(\alpha_{ij} = \alpha)$ is obtained by averaging the probability assigned to α by $p(\cdot|\theta)$, for all $\theta \in \Theta$,

$$q(\alpha) = \int_{\Theta} p(\alpha|\theta)w(\theta)d\theta, \tag{5}$$

where $w(\theta)$ is an (hyper-)prior on θ. In Bayesian theory, $w(\theta)$ reflects the prior belief on the values of θ. This is the main idea behind sparse Bayesian coding works such as [14, 29]. However, in universal coding/MDL theory such interpretation is not necessary, and it can be shown that any *smooth* choice $w(\theta)$ is enough to guarantee the universality of the resulting mixture [2]. This allows us to obtain a closed form solution of (5) for Laplacian $p(\cdot|\theta)$ in the following way. Since the Laplacian is a symmetrized version of the Exponential distribution, we can perform the mixture on the Exponential distribution and then add back the symmetry by substituting α with $|\alpha|$ and dividing the normalization constant by $1/2$. A closed form solution of (5) can be obtained by using the conjugate prior for the Exponential, which is the Gamma distribution,

$$w(\theta|\kappa, \beta) = \Gamma(\kappa)^{-1}\theta^{\kappa-1}\beta^{\kappa}e^{-\beta\theta},$$

where κ and β are its *shape* and *scale* parameters respectively. By using the Gamma prior, and later re-symmetrizing the distribution we arrive at the following symmetric distribution,

$$q_{\text{MOL}}(\alpha|\beta, \kappa) = \kappa\beta^{\kappa}(|\alpha| + \beta)^{-(\kappa+1)}, \tag{6}$$

which we call a *Mixture of Laplacians* (MOL). Although the resulting prior has two parameters to deal with instead of one, it will be shown later in Section 7.1, that a single MOL distribution can fit each of the K rows of $\mathbf{A}$ better than K separate Laplacian distributions fine-tuned to these rows, for a total of K parameters to be estimated. Furthermore, both κ and β are easily computed using the method of moments. It is easy to see that the non central moment of order j of the MOL distribution is $\mu_j = \beta^j / \binom{\kappa-1}{j}$. Thus, given sample estimates $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}|\alpha|$ and $\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}|\alpha|^2$ we have

$$\hat{\kappa} = 2(\hat{\mu}_2 - \hat{\mu}_1^2)/(\hat{\mu}_2 - 2\hat{\mu}_1^2) \quad \text{and} \quad \hat{\beta} = (\hat{\kappa} - 1)\hat{\mu}_1, \tag{7}$$

When the MOL prior is plugged into (4), the following new cost function is obtained,

$$f(\mathbf{X}, \mathbf{A}|\mathbf{D}) = \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \tau \sum_{j=1}^{N}\sum_{i=1}^{K} \log\left(|\alpha_{ij}| + \beta\right), \tag{8}$$

where $\tau = 2\sigma^2(\kappa + 1)$. The resulting logarithmic non-convex MOL regularization term, $\log p(\mathbf{A}|\mathbf{D})$, is known in robust statistics as the *Lorentzian* norm, also known to be more robust to outliers than the ℓ_1 norm. We also know from the statistics literature that the MOL regularizaton term leads to consistent estimators of regression coefficients which are able to identify the relevant variables in a regression model (oracle property) [13]. This is not the case for the ℓ_1 regularizer [31]. This same regularizer has also been recently proposed in the context of compressive sensing [5], where it is conjectured to be better than the ℓ_1-term at recovering sparse signals.[1] Our results in Section 7 give evidence that this is indeed the case, with the direct consequence of a much improved reconstruction accuracy of sparse data. We also show in Section 7 that the MOL prior is much better to model reconstruction coefficients drawn from a large database of image patches. We will also see next that although the MOL regularizer is non-convex, simple and effective methods are available to solve the resulting sparse coding (or regression) problems.

Remark: Instead of using a Gamma prior (which actually already fits the empirical statistics of image data very good [26]), we can use for example the Jeffreys prior, which has several interesting properties both from the information-theoretic and statistical points of view. The Jeffreys prior for a given distribution is defined as

$$w_J(\theta) = \frac{\sqrt{I(\theta)}}{\int_{\Theta}\sqrt{I(\theta)}d\theta}$$

[1] In [5], the logarithmic regularizer arises from approximating the ℓ_0 pseudo-norm as a ℓ_1-normalized element-wise sum.

where $I(\theta)$ is the *Fisher Information matrix* of a parametric distribution of parameter θ:

$$I(\theta) = \left\{ E_{p(\cdot|\tilde{\theta})} \left[-\frac{\partial^2}{\partial \tilde{\theta}^2} \log p(x|\tilde{\theta}) \right] \right\} \Big|_{\tilde{\theta}=\theta} . \tag{9}$$

For the Exponential distribution (which is the one-sided version of the Laplacian) we have that $I(\theta) = \frac{1}{\theta^2}$. Clearly, if we let $\Theta = (0, \infty)$, the integral in (9) evaluates to ∞. To get a proper integral, we restrict Θ to be a closed interval $[a, b]$, $0 < a < b < \infty$. For this choice we get $w_J(\theta) = \frac{1}{\ln(b/a)} \frac{1}{\theta}$. The resulting mixture, after being symmetrized around 0, has the following form

$$q_{\text{JOL}}(\alpha) = \frac{1}{\ln(b/a)} \frac{1}{|\alpha|} \left(e^{-a|\alpha|} - e^{-b|\alpha|} \right) . \tag{10}$$

We refer to this prior as a Jeffreys mixture of Laplacians (JOL). Note that although q_{JOL} is not defined for $\alpha = 0$, its limit when $\alpha \to 0$ is finite and evaluates to $\frac{b-a}{2\ln(b/a)}$. Thus, by defining $q_{\text{JOL}}(0) = \frac{b-a}{2\ln(b/a)}$ we obtain a prior that is well defined and continuous for all $\alpha \in \mathbb{R}$. Experimental results with this prior, as well as the full mathematical details of the derivations here described are provided in [27].

5 Proposed sparse model

To the model in (8), which replaces the more classical (1), we add the term for dictionary incoherence introduced in Section 3, leading to the proposed sparse model

$$f(\mathbf{X}, \mathbf{D}, \mathbf{A}) = \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \tau \sum_{j=1}^{N} \sum_{i=1}^{K} \log\left(|\alpha_{ij}| + \beta\right) + \zeta \left\|\mathbf{D}^T\mathbf{D} - \mathbf{I}_K\right\|_F^2 + \eta \sum_{k=1}^{K} (\|\mathbf{D}_k\|_2^2 - 1)^2 . \tag{11}$$

The last term is added as a standard way to maintain the atom norms close to one.[2] This (soft) normalization was empirically observed to yield better results than a forced projection of the atoms after each update, which is the approach followed in [1, 12, 24].

6 Numerical optimization

To minimize (11) we apply the standard approach of alternate minimization. We start with an arbitrary initial dictionary $\mathbf{D}^{(0)}$ and repeat the following sequence of updates until convergence

$$\text{SC}: \mathbf{A}^{(t+1)} = \arg\min_{\mathbf{A}} \left\{ f(\mathbf{X}, \mathbf{D}^{(t)}, \mathbf{A}) \right\} \text{ and } \text{DU}: \mathbf{D}^{(t+1)} = \arg\min_{\mathbf{D}} \left\{ f(\mathbf{X}, \mathbf{D}, \mathbf{A}^{(t+1)}) \right\} . \tag{12}$$

Sparse coding (SC) For fixed $\mathbf{D}$, the cost function in (11) is non-convex in $\mathbf{A}$. We handle this issue as in [5], using a *surrogate function* technique where the optimal solution $\mathbf{A}^{(t+1)}$ to the SC problem is the limit of solutions to a sequence of subproblems, where the non-convex cost function is approximated by convex functions that are easy to solve. More specifically, we have that $\mathbf{A}^{(t+1)} = \lim_{z\to\infty}\{\Psi^{(z)}\}$, with $\Psi^{(z)} = \arg\min_{\Psi} \left\{ h^{(z)}(\mathbf{X}, \mathbf{D}, \Psi) \right\}$.

[2] Note that the term $\left\|\mathbf{D}^T\mathbf{D} - \mathbf{I}_K\right\|_F^2$ does not enforce the atoms to be normalized.

The surrogate $h^{(z)}(\cdot)$ is obtained by a first order expansion of the logarithmic terms in $f(\cdot)$ around the previous iterate $\Psi^{(z-1)}$,

$$\log(|\psi| + \beta) \leq \log(|\psi^{(z-1)}| + \beta) + \frac{|\psi| - |\psi^{(z-1)}|}{|\psi^{(z-1)}| + \beta},$$

where ψ denotes a generic element of the matrix Ψ. This optimization technique is a special case of the Local Linear Approximation (LLA) technique described in [32], where it was shown to converge to a stationary point. Discarding the constant terms and defining $a_{ij}^{(z-1)} = (|\psi_{ij}^{(z-1)}| + \beta)^{-1}$, we get the following sequence of subproblems:

$$\Psi^{(z)} = \arg\min_{\Psi} \left\{ \|\mathbf{X} - \mathbf{D}\Psi\|_2^2 + \tau \sum_{j=1}^{N} \sum_{i=1}^{K} a_{ij}^{(z-1)} \, |\psi_{ij}| \right\}. \tag{13}$$

We set $\psi_{ij}^{(0)} = 0$, so that the first iterate $\Psi^{(1)}$ is the solution to the *unmixed* model (1) with $\lambda = 2\sigma^2(\kappa+1)/\beta$, which we expect to be a good starting point for the minimization. For $z > 1$, the problem in (13) corresponds to a weighted version of (1), which can be solved with any ℓ_1 solver.

Dictionary update (du) Given a current estimation for $\mathbf{A}$, a popular choice for the dictionary update step, used in many current state of the art applications (e.g. [21]), is the Method of Optimal Directions (MOD) [12] (K-SVD could be similarly used). The MOD updates $\mathbf{D}$ using the standard least squares estimator $\mathbf{D}^{(t+1)} = \mathbf{X}(\mathbf{A}^{(t+1)})^T(\mathbf{A}^{(t+1)}(\mathbf{A}^{(t+1)})^T)^{-1}$. In our case, the updated dictionary $\mathbf{D}^*$ is obtained by taking the derivative $\nabla_{\mathbf{D}} f(\mathbf{D})$ of (11) with respect to $\mathbf{D}$, and solving $\nabla_{\mathbf{D}} f(\mathbf{D}) = 0$ for $\mathbf{D}$. The resulting update is called MOCOD (Method of Optimal COherence-COnstrained Directions),

$$\begin{aligned} \mathbf{D}^{(t+1)} &= \left(\mathbf{X}(\mathbf{A}^{(t+1)})^T + 2(\zeta+\eta)\mathbf{D}^{(t)}\right) \times \\ &\quad \left[\mathbf{A}^{(t+1)}(\mathbf{A}^{(t+1)})^T + 2\,\zeta\,(\mathbf{D}^{(t)})^T\mathbf{D}^{(t)} + 2\,\eta\,\mathrm{diag}\left((\mathbf{D}^{(t)})^T\mathbf{D}^{(t)}\right)\right]^{-1}. \end{aligned} \tag{14}$$

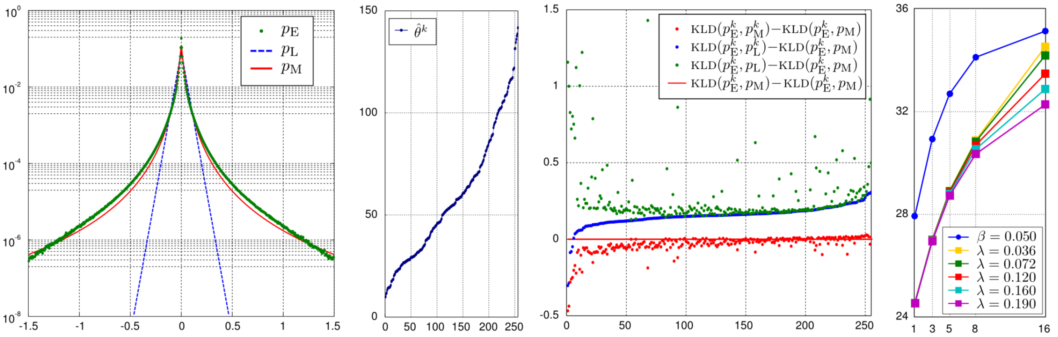
Note that when $\zeta = 0$ and $\eta = 0$ the MOCOD update (14) coincides with MOD.

7 Experimental results: Sparse coding using mol

In the following experiments, all images are converted to grayscale with an intensity range of $[0,1]$ and then broken into patches of size 8×8 pixels, yielding data vectors of dimension $n = 64$. Similar results to those shown here are also obtained for other patch sizes. We choose not to include them due to space constrains.

The first set of experiments studies the properties of the MOL distribution for sparse coding only. Its properties with respect to the whole dictionary learning model will be discussed in Section 8 along with the other added terms. For these experiments we use a global dictionary of $K = 256$ atoms trained to the Pascal VOC2006 *training* subset[3] using the model (1) with $\lambda = 0.1$. These parameter values are typical in sparse coding applications and produce dictionaries $\mathbf{D}$ that lead to state-of-the-art results [1, 20, 22].

[3] http://pascallin.ecs.soton.ac.uk/challenges/VOC/databases.html#VOC2006



(a) (b) (c) (d)

Figure 1: (a) Empirical distribution of reconstruction coefficients for image patches and the best fitting distributions. (b) Variation of the Laplacian parameter $\hat{\theta}^k$ in p^k_{L} for all atoms (sorted in ascending $\hat{\theta}^k$). (c) Differences between the Kullback-Leibler Divergences for the best fitting distributions computed per atom (see Section 7.1 for details). (d) Reconstruction PSNR for the proposed MOL and classical ℓ_1 formulations for different ℓ_0 (see Section 7.3 for details).

7.1 MOL as a prior for reconstruction coefficients

We now evaluate the goodness of fit of the Laplacian and MOL distributions to the empirical distribution of the reconstruction coefficients $\mathbf{A}$ obtained by considering all its elements $\{\alpha_{kj}\}$ as IID. We compute $\mathbf{A}$ using BP to obtain an exact reconstruction,

$$\min\left\{\|\mathbf{A}\|_1 \,\text{s.t.}\, \mathbf{X} = \mathbf{DA}\right\},$$

and then restrict our study to the nonzero elements of $\mathbf{A}$. $\mathbf{X}$ corresponds to all 8×8 patches from the 2600 *testing* images from the Pascal VOC2006 dataset.

The empirical distribution of $\mathbf{A}$, p_{E}, is plotted in Figure 1a (green dots) along with the best fitting Laplacian, p_{L}, (in blue) and MOL, p_{M}, distributions (in red). The fitting is done using the MLE estimator for the Laplacian parameter, $\hat{\theta} = N/\left\|\mathbf{A}\right\|_1$, and for MOL we fix $\kappa = 3$ and compute β using the method of moments, which gives $\hat{\beta} = (\kappa - 1)\left\|\mathbf{A}\right\|_1/N$. The best fitting distributions were obtained for $\hat{\theta} = 55$ and $\hat{\beta} = 0.05$ respectively. Visual inspection of Figure 1a reveals a much better fitting of MOL to the empirical distribution. For an objective measure of the goodness of fit we use the standard Kullback-Leibler Divergence (KLD) between p_{E} and each of the fitted distributions p_{L} and p_{M}, yielding $\mathrm{KLD}(p_{\mathrm{E}}, p_{\mathrm{L}}) = 0.30$ bits and $\mathrm{KLD}(p_{\mathrm{E}}, p_{\mathrm{M}}) = 0.04$, confirming the improved fitting properties of MOL. The empirical entropy is $H(p_{\mathrm{E}}) = 3.00$ bits.

Figure 1b shows that the optimal Laplacian parameter $\hat{\theta}^k$ varies greatly with each atom (sorted in ascending $\hat{\theta}^k$). This justifies the following experiment, where different Laplacian and MOL distributions are fitted to each k row of $\mathbf{A}$, $\mathbf{A}^k$, each corresponding to the k-th atom in $\mathbf{D}$. We then compare the fitting obtained by the row-specific empirical distributions, p^k_{E}, against the row-specific fitted Laplacian, p^k_{L}, and MOL, p^k_{M}, and against the globally-fitted ones p_{L} and MOL p_{M}. Figure 1c plots the difference between the KLDs for the computed distributions and $\mathrm{KLD}(p^k_{\mathrm{E}}, p_{\mathrm{M}})$, with the horizontal axis sorted by increasing $\mathrm{KLD}(p^k_{\mathrm{E}}, p^k_{\mathrm{L}}) - \mathrm{KLD}(p^k_{\mathrm{E}}, p_{\mathrm{M}})$. Clearly, the distributions of the same type (Laplacian or MOL) fitted to each row should be at least as accurate for the statistics of that row as the globally fitted ones, and this can be verified in Figure 1c. We then observe that p_{M} is significantly better than

p_{L} for every k. Furthermore, we observe that in 251 out of 256 cases the *global* p_{M} performs better than the row-specific p_{L}^k. Finally, in all cases, the row-specific p_{M}^k are better than the corresponding p_{L}^k by an approximately constant, significant margin.

We conclude that MOL, with the same number of free parameters than a Laplacian, is significantly better for the probabilistic modeling of reconstruction coefficients. This also holds for row-wise elements, providing further evidence that the univeral approach is a suitable strategy for handling the non-stationary nature of image patch statistics. Finally, the fact that a single globally fitted MOL distribution performs better, in almost all cases, than K Laplacians fitted specifically to each row, shows that our approach is also more accurate, with only one parameter, than a weighted ℓ_1 model with K parameters.

Wether these significantly improved fittings have a practical impact on the performance of the sparse coding stage is explored in the following experiments.

7.2 Active set recovery

For the following experiments we consider data, available as part of the SIPI database,[4] consisting of the six widely used grayscale images *Barbara*, *Boats*, *Baboon*, *Goldhill*, *Lena* and *Peppers*. We perform an empirical study on the active set recovery properties of the MOL prior compared to those of the ℓ_1-based one. To this end, we first decompose the images of the dataset into non-overlapping 8×8 patches and obtain sparse approximations to them using OMP for different considered target sparsity levels ℓ_0. This way we have a set of signals which have a true sparse representation under the dictionary $\mathbf{D}$, and whose active set can be used as a ground truth. For each target maximum ℓ_0 norm we do an exact reconstruction of the patches using BP (corresponding to the ℓ_1 cost) and the "BP equivalent" for the MOL prior,

$$\mathbf{A}^* = \arg\min_{\mathbf{A}} \left\{ \sum_{j=1}^{N} \sum_{i=1}^{K} \log\left(|\alpha_{ij}| + \beta\right) \;\text{ s.t. }\; \mathbf{X} = \mathbf{DA} \right\}.$$

We then measure the accuracy of the recovery of each method as the percentage of cases $\mathcal{H}(n_\varepsilon)$ in which the size of the symmetric difference between the true and recovered active sets, $|(A \setminus B) \cup (B \setminus A)|$, is no larger than a certain target value n_ε. In order to quantify how this accuracy relates to the final reconstruction quality of the patches, we measure the reconstruction PSNR obtained using an Ordinary Least Squares (OLS) approximation restricted to the recovered active sets (this is of course the appropriate reconstruction procedure once the active set is determined).

Table 1a shows these results for different maximum ℓ_0 values. As can be observed, in all cases the MOL-based recovery is significantly more successful in recovering the true active sets, and this translates into a much better reconstruction performance.

7.3 Sparse coding performance of MOL

Given the previous results, we expect sparse coding based on the MOL prior (8) to outperform the one obtained using an ℓ_1 prior (1). Clearly both priors have different forms, whose parameters play different roles, and thus comparing the reconstruction PSNR of both formulations needs to be done by carefully matching the regularization strength of the ℓ_1

[4] http://sipi.usc.edu/database/

and MOL terms. Since MOL is a mixture of Laplacians, one such way is to fix the same value of σ^2 in both formulations and let $\lambda = 2\sigma^2\bar{\theta}$ where $\bar{\theta} = E[\theta]$ is the expected value of θ under the Gamma distribution used as a prior on θ for computing the MOL distribution.

For this case we encode the SIPI subset by solving both ℓ_1 and MOL-based sparse coding problems. We use $\kappa = 3$ and the optimal scale parameter $\beta = 0.05$ found in Section 7.1 for (8) and a range of values of $\theta = \{0.3, 0.6, 1.3, 1.6\}\,\bar{\theta}$ where $\bar{\theta} = \kappa/\beta$ is the expected value of θ corresponding to β, as discussed in Section 4. In both cases we set $\sigma = 3\%$ of the peak value in order to define λ in (1) and τ in (8). In order to better observe the differences in performance, we truncate the solutions to have a fixed maximum ℓ_0 norm and compute the reconstruction using OLS based on the columns of $\mathbf{D}$ chosen by the active set only. The results in Figure 1d show that we indeed have a better sparse reconstruction with a single value of β than *any* value of λ, thus confirming the advantages of using MOL in this case as well.

8 Experimental results: Testing mocod

We now show that the complete proposed model (Section 5) improves both the reconstruction accuracy and the speed of leading sparse coding techniques.

8.1 Reconstruction and generalization properties

The following experiment deals with the generalization properties of globally trained dictionaries. In this case we expect that the designed incoherence, which can be seen as a constrain on how close the different learned atoms are in $\mathbb{R}^K$, reduces overfitting to the training data by avoiding the clumping of atoms in regions where there are exceptional concentrations of similar training vectors.

We take the images from the SIPI subset and perform a *leave-one-out* cross-validation procedure where each image is encoded using a dictionary trained with the other images. For training, a small amount of white Gaussian noise with standard deviation $\sigma = 2\%$ of the peak pixel value is added so that the parameter σ^2 that defines λ in (1) and τ in (8) is known. The initial dictionary is the one used in Section 7. After the dictionary is learned, the patches of the target image are approximated using OMP with a maximum of $\ell_0 = 12$ nonzero reconstruction coefficients per patch. The image patches in the reconstruction step are non-overlapping.[5] The MOL parameters were set to $\kappa = 3$ and the global optimum found in Section 7.1 $\beta = 0.05$, and $\theta = \{0.5, 1.0, 1.6\}\,\bar{\theta}$, where $\bar{\theta} = \kappa/\beta$.

The results are shown in Table 1b, where we compare MOCOD against MOD, showing that dictionaries trained with MOCOD yield a significant average improvement of up to 1 dB in reconstruction PSNR, thus providing evidence of the improved generalization properties. In all cases we see that the Gram matrix norm $\rho(\mathbf{D})$ and the cumulative coherence $\bar{\mu}(\mathbf{D})$ are significantly reduced as expected. We omit a comparison to K-SVD [1] since it was observed in [21] that both MOD and K-SVD give very similar results in general.

[5] For reconstruction, specially for denoising, state-of-the art algorithms rely on overlapping. We use non-overlapping patches in these experiment so that differences in the quality of the reconstructed patches are not averaged out.

ℓ_0	n_ε	SC	$\mathcal{H}(n_\varepsilon)$	OLS
				PSNR
3	0	ℓ_1	35.6	37.4
		MOL	**71.1**	**42.6**
5	1	ℓ_1	10.6	36.9
		MOL	**43.2**	**42.2**
8	2	ℓ_1	7.6	37.6
		MOL	**30.8**	**42.3**

(a)

DU	SC	ρ	$\bar{\mu}$	OMP	IST	
				PSNR	t (sec.)	PSNR
MOD	ℓ_1 $(\theta = 150)$	15.8	39.5	38.7	8.3	33.8
MOCOD		**7.5**	**24.9**	**39.1**	**4.9**	**34.0**
MOD	ℓ_1 $(\theta = 60)$	11.5	31.3	38.0	7.2	33.6
MOCOD		**7.1**	**24.7**	**39.0**	**4.7**	**33.9**
MOD	ℓ_1 $(\theta = 30)$	15.1	36.5	38.4	8.1	33.8
MOCOD		**6.8**	**24.4**	**38.9**	**4.5**	33.8
MOD	MOL $(\beta = 0.05)$	20.0	43.2	38.6	9.7	33.8
MOCOD		**7.0**	**24.6**	**39.0**	**4.8**	**33.9**

(b)

Table 1: (a) Accuracy of the recovery of active set and its PSNR (see text, Section 7.2). (b) Reconstruction properties. The best results for each case are in bold. The values of β, θ, ρ and $\bar{\mu}$ and the OMP PSNR (in dB) columns correspond to the experiment of Section 8.1, while the last two columns are for the IST performance results of Section 8.2.

8.2 Improved computational efficiency of shrinkage methods and active set recovery

As mentioned in Section 3, the convergence rate of *Iterated Shrinkage/Thresholding* [7], and of state-of-the art sparse coding methods based on it (see [11] for a review), are known to depend on the Gram matrix norm $\rho(\mathbf{D})$. Since MOCOD induces a reduction in $\rho(\mathbf{D})$, it is expected that dictionaries learned with MOCOD lead to a faster convergence than the ones obtained without the additional coherence penalty. The last two columns in Table 1b account for the running time and final PSNR obtained using IST for a fixed value of $\lambda = 0.1$ and the corresponding $\mathbf{D}$.[6] The coding time obtained with the dictionaries trained with MOCOD is consistently reduced by an amount roughly proportional to the reduction in $\rho(\mathbf{D})$, while the PSNR remains approximately the same (or improves slightly).

Finally, we repeated the experiment in Section 7.2 using the complete MOCOD and observed further significant improvements of the values in the two rightmost columns in Figure 1a, for example, an improvement of 61% in $\mathcal{H}(n_\varepsilon)$ for $\ell_0 = 8$, $n_\varepsilon = 2$, and of 2.7dB in PSNR for $\ell_0 = 3$. This being a direct consequence of reducing $\bar{\mu}(\mathbf{D})$, it further shows that the incoherence term, on top of MOL, is very important for efficient and accurate reconstruction of both the active set and the signal itself.

9 Concluding remarks

A new sparse model was introduced in this work. The model includes two new terms. The first new component encourages intrinsic properties of the learning dictionary which are critical for sparse coding and generalization. The second term replaces the classical sparsifying penalties by a logarithmic one formally derived following the framework of MDL. Both the theoretical and practical advantages of such new model were presented.

The critical properties of the proposed model, such as increased stability of the active set and improved generalization, hint to the possible implications of this model for classification

[6] Timings were obtained using a single-threaded C implementation compiled with GCC 4.3.2, on a Lenovo T400 with Core2 Duo T9400 (at full speed) running Linux 2.6.27.

tasks such as those described in [20]. We are currently investigating this and results, together with the theoretical details on the derivations here presented, will be reported elsewhere.

Acknowledgements

Work partially supported by ONR, NGA, NSF, ARO, DARPA and FUNDACIBA-ANTEL. We thank Julien Mairal for his help with software and his very illuminating insights about this work.

References

[1] M. Aharon, M. Elad, and A. M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Trans. SP*, 54(11):4311–4322, Nov. 2006.

[2] A. R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. IT*, 44(6):2743–2760, 1998.

[3] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, Feb. 2009.

[4] E. J. Candès, J. K. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. IT*, 52(2):489–509, 2006.

[5] E. J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.*, 14(5):877–905, Dec. 2008.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[7] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. on Pure and Applied Mathematics*, 57:1413–1457, 2004.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[9] M. Elad. Optimized projections for compressed-sensing. *IEEE Trans. SP*, 55(12):5695–5702, Dec. 2007.

[10] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. IP*, 54(12):3736–3745, Dec. 2006.

[11] M. Elad, B. Matalon, J. Shtok, , and M. Zibulevsky. A wide-angle view at iterated shrinkage algorithms. In *SPIE (Wavelet XII)*, pages 26–29, San-Diego CA, Aug. 2007.

[12] K. Engan, S. O. Aase, and J. H. Husoy. Multi-frame compression: Theory and design. *Signal Processing*, 80(10):2121–2140, Oct. 2000.

[13] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal Am. Stat. Assoc.*, 96(456):1348–1360, Dec. 2001.

[14] M. A. T. Figueiredo. Adaptive sparseness using Jeffreys prior. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Adv. NIPS*, pages 697–704. MIT Press, 2001.

[15] R. Giryes, Y. C. Eldar, and M. Elad. Automatic parameter setting for iterative shrinkage methods. In *IEEE 25-th Convention of Electronics and Electrical Engineers in Israel (IEEEI'08)*, Dec. 2008.

[16] G. H. Golub and C. F. van Loan. *Matrix Computations.* JHU Press, 3rd edition, 1996.

[17] S. D. Howard, A. R. Calderbank, and S. J. Searle. A fast reconstruction algortihm for deterministic compressive sensing using second order Reed-Muller codes. In *Conf. on Info. Sciences and Systems (CISS), Princeton, New Jersey*, Mar. 2008.

[18] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. PAMI*, 27(6):957–968, 2005.

[19] M. S. Lewicki and B. A. Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A*, 16(7):1587–1601, 1999.

[20] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Adv. NIPS*, volume 21, 2009.

[21] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce. Discriminative sparse image models for class-specific edge detection and image interpretation. In *Proc. ECCV*, Oct. 2008.

[22] J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *SIAM MMS*, 7(1):214–241, Apr. 2008.

[23] S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Trans. SP*, 41(12):3397–3415, 1993.

[24] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[25] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML*, pages 759–766, 2007.

[26] I. Ramirez, F. Lecumberry, , and G. Sapiro. Universal priors for sparse modeling. *IMA Preprint, http://www.ima.umn.edu/preprints/aug2009/2276.pdf*, 2009.

[27] I. Ramirez, F. Lecumberry, and G. Sapiro. Universal priors for sparse signal modeling: Marrying information theory with sparse coding. 2009, in preparation.

[28] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *IEEE Trans. SP*, 56(5):1994–2002, 2008.

[29] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[30] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. IT*, 50(10):2231–2242, Oct. 2004.

[31] H. Zou. The adaptive LASSO and its oracle properties. *Journal Am. Stat. Assoc.*, 101:1418–1429, 2006.

[32] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.