UNIVERSAL PRIORS FOR SPARSE MODELING

By

Ignacio Ramírez
Federico Lecumberry
and
Guillermo Sapiro

IMA Preprint Series # 2276
(August 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
# Universal Priors for Sparse Modeling

**Title:** Universal Priors for Sparse Modeling  
**Report Date:** AUG 2009  
**Dates Covered:** 00-00-2009 to 00-00-2009

**Performing Organization:**  
University of Minnesota, Institute for Mathematics and Its Applications, Minneapolis, MN, 55455-0436

**Distribution/Availability Statement:**  
Approved for public release; distribution unlimited

**Security Classification:**  
- Report: Unclassified  
- Abstract: Unclassified  
- This Page: Unclassified 

**Limitation of Abstract:** Same as Report (SAR)

**Number of Pages:** 5
Universal Priors for Sparse Modeling

(Invited Paper)

Ignacio Ramírez#1, Federico Lecumberry#2, Guillermo Sapiro#3

#Electrical Engineering Department, University of Minnesota
200 Union Street S.E., MN 55455, USA
1,3{ramir048,guille}@umn.edu

#Instituto de Ingeniería Eléctrica, Universidad de la República
J. Herrera y Reissig 565, Montevideo 11300, Uruguay
2{fefo@fing.edu.uy}

Abstract—Sparse data models, where data is assumed to be well represented as a linear combination of a few elements from a dictionary, 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. It is now well understood that the beneficial effect of the sparsity regularization term is critical in the success of such models. In this work, we use tools from information theory to propose a sparsity regularization term that leads to a better sparse reconstruction than ones, and which leads to improved coding performance and accuracy in reconstruction tasks. We also briefly report on further improvements obtained by imposing low mutual coherence and Gram matrix norm on the learned dictionaries.

I. 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, e.g., [1], [12], [20], and of representing the actual data in terms of them, e.g., [6], [8], [9], have been developed in recent years, leading to state-of-the-art results in many signal and image processing tasks [11], [16], [17]. We refer the reader for example to [3] for a recent review on the subject.

A critical component of sparse modeling is the actual sparsity of the representation, which is controlled by some 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], [24]. In this paper we address this challenge, and at the same time further generalize the standard sparsifying penalty functions (or priors for short), exploiting tools from information theory. The result is a prior that 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 \( \ell_0 \) and \( \ell_1 \)-based techniques in practice. This new model is complemented by imposing incoherence in the learned dictionary.

II. SPARSE MODELING

Let \( X \in \mathbb{R}^{M \times N} \) be a set of \( N \) column data samples \( X_j \in \mathbb{R}^M \), \( D \in \mathbb{R}^{M \times K} \) be a dictionary of \( K \) atoms represented as columns \( D_k \in \mathbb{R}^M \), and \( A = (a_{kj}) \in \mathbb{R}^{K \times N} \). We use an approach similar to [3] to find an approximate solution is alternate minimization. Starting with an initial dictionary \( D^{(0)} \), the following sequence of subproblems is repeated until convergence,

\[
(A^{(t+1)}, D^{(t+1)}) = \arg \min_{A, D} \| X - DA \|_F^2 + \lambda \| A \|_1
\]

where \( \| \cdot \|_F \) denotes Frobenius norm. The \( \ell_1 \) norm is used as an approximation to \( \ell_0 \), making the problem convex in \( A \) while still encouraging sparse solutions [3].

III. UNIVERSAL MODELS FOR SPARSE CODING

Given a fixed dictionary \( D \), the problem of finding the coefficients \( A \) that minimizes (1) can be viewed as a Maximum a Posteriori (MAP) estimation in the logarithmic scale, that is

\[
A^* = \max_A \log p(X|A) + \log p(A),
\]

where \( p(X|A) \propto \exp(-\frac{1}{2\sigma^2} \| X - DA \|_2^2) \), and the prior on \( A \) is \( \text{IID} \) Laplacian with mean 0 and inverse-scale parameter \( \theta \), \( p(A) \propto \exp(-\theta \| A \|_1) \). The energy term in Equation (1) follows by taking the logarithms of both priors and factorizing 2\( \sigma^2 \) into \( \lambda = 2\sigma^2\theta \).
Even for the Laplacian IID model, the problem of finding the optimal parameter $\lambda$ (or $\theta$, for given $\sigma^2$) is already a challenging problem (see for example [15]).

In this work we consider an independent (but not identically distributed) Laplacian model where the underlying parameter $\theta$ can be different for each atom $k$ and, furthermore, where each of these $\theta^k$ can also vary across samples. This scenario is justified when modeling small patches from natural images, which is our primary type of data.

Assuming a known parametric form for the prior $\theta$ leads to the concept of a model class. In our case, we consider the class $\mathcal{M} = \{p(\cdot|\theta): \theta \in \Theta\}$ of all Laplacian models $p(\cdot|\theta)$ with $\theta \in \Theta \subseteq \mathbb{R}^+ / \{0\}$. The goal now is to find a probability model for $\Lambda$ which can fit each $\Lambda_j$ as well as the model in $\mathcal{M}$ that can be fitted to $\Lambda_j$, after having observed it, for every sample $j = 1, \ldots, N$. The construction of such universal models (meaning that they are universally good with respect to any model from $\mathcal{M}$) is subject to the universal coding theory, which lies at the core of the Minimum Description Length principle (MDL) [2].

We now briefly discuss the principles of universal coding, and how they apply in our case. In the following discussion we consider the reconstruction coefficients data to be a one-dimensional sequence of $n$ scalar values $\alpha^0 = (\alpha_1, \ldots, \alpha_n)$.

For given data $\alpha^n$ we measure the goodness of fit of a model $q(\cdot)$ with respect to $\mathcal{M}$ using the code length regret,

$$R(\alpha^n, q) := -\log q(\alpha^n) + \log p(\alpha^n | \hat{\theta}(\alpha^n)),$$

where $\hat{\theta}(\alpha^n)$ is the Maximum Likelihood Estimator of $\theta$ for the observed $\alpha^n$. Minimizing the worst-case regret,

$$q^* = \min_q \max_{\alpha^n} -\log q(\alpha^n) + \log p(\alpha^n | \hat{\theta}(\alpha^n)),$$

leads to the Normalized Maximum Likelihood (NML) distribution, $q^*(\alpha^n) = \log p(\alpha^n | \hat{\theta}(\alpha^n))/C(\mathcal{M}, n)$. The normalization constant $C(\mathcal{M}, n) = \int_\Theta p(\alpha^n | \theta) d\theta$ is also the value of the minimax regret and depends only on $\mathcal{M}$ and the length of the data $n$. Since the regret of the NML code is at most $C(\mathcal{M}, n)$ for any sequence $\alpha^n$, the NML model $q^*$ is said to be worst case universal.

Unfortunately, the NML model depends on the observed data, so it is not suitable to be used as a prior for computing the data itself. However, we can obtain a different universal model, which can be used as a prior, by considering instead the expected regret. This is defined with respect to a given distribution $p(\cdot | \theta)$ as

$$R(p(\cdot | \theta), q) = E_{p(\cdot | \theta)}[-\log q(\alpha^n) + \log p(\alpha^n | \hat{\theta}(\alpha^n))].$$

The expected regret can be further averaged with respect to some hyper-prior on $\theta$ itself, $w(\theta)$. It is straightforward to see that this is equivalent to computing the expected regret with respect to the Bayes mixture

$$q^w(\cdot) = \int_{\Theta} p(\alpha^n | \theta) w(\theta) d\theta. \quad (3)$$

Since the regret in this case equals

$$R(q^w, q) = E_{q^w}[-\log q^w(\alpha^n) + \log p(\alpha^n | \hat{\theta})] + D(q^w||q),$$

where $D(\cdot||\cdot)$ is the Kullback-Leibler divergence (KLD), it is also trivial to see that $R(q^w, q)$ is minimized for $q = q^w$.

An important result in universal coding and MDL theory is that, for smooth parametric families such as the Laplacian, the worst case regret of the Bayes mixture obtained for any smooth choice of $w(\theta)$ is within $O(1)$ of the NML regret [2]. This allows us to choose a prior that is computationally practical, such as the conjugate prior for the Laplacian, which is the Gamma distribution, $w(\theta|\kappa, \beta) = \Gamma(\kappa)^{-1}\theta^{\kappa-1}\beta^\kappa e^{-\beta\theta}$. Here $\kappa$ and $\beta$ are the shape and scale parameters of the Gamma distribution respectively. Note that, in Bayesian theory, $w(\theta)$ reflects the prior belief on the values of $\theta$. This is the main idea behind sparse Bayesian coding works such as [14], [22]. As mentioned, results in universal coding [2] tell us that it is good enough for $w(\theta)$ to be smooth to obtain a code that is (asymptotically) good. However, quite strikingly, it was observed in practice that the Gamma distribution is indeed very good for modeling spatial variations in the optimal value of $\theta$ (see Figure 1c along with the discussion in Section IV).

Plugging the Laplacian as $p(\cdot|\theta)$ and the Gamma prior as $w(\theta)$ into (3) results in

$$q(\alpha|\beta, \kappa) = 0.5\kappa e^{\beta(\alpha/\beta - \kappa + 1)}, \quad (4)$$

which we call a Mixture of Laplacians (MOL). The parameters can be estimated from data using the method of moments as

$$\kappa = 2(\mu_2 - \mu_1^2)/\mu_2^2 \quad \text{and} \quad \beta = (\kappa - 1)\mu_1, \quad (5)$$

where $\mu_j = \sum_{i=1}^n |x_i|^j$ are the $j$-th non-central absolute sample moments. The role of $\beta$ is equivalent to $1/\theta$, that is, it controls the scale of the prior. When the MOL prior (4) is plugged into (2), the resulting MAP sparse coding model is

$$A^* = \arg \min_\Lambda \|X - DA\|^2_F + \tau \sum_{j=1}^N \sum_{k=1}^K \log (|\alpha_{kj}| + \beta),$$

where $\tau = 2\sigma^2(\kappa + 1)$. The resulting logarithmic non-convex MOL regularization term is known in robust statistics as the Lorentzian norm, also known to be more robust to outliers than the $\ell_1$ norm. We also know from the statistics literature that the MOL regularization 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 $\ell_1$ regularizer [24]. This same regularizer has also been recently proposed in the context of compressive sensing [5], where it is conjectured to be better than the $\ell_1$-term at recovering sparse signals. Our results in Section IV 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 IV 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.

Finally, we combine the new prior with two additional terms that apply to the dictionary D into the following formulation

\[
(\mathbf{A}^*, \mathbf{D}^*) = \arg \min_{\mathbf{A}, \mathbf{D}} \| \mathbf{X} - \mathbf{D}\mathbf{A} \|_F^2 + \gamma \sum_{j=1}^N \sum_{k=1}^K \log (|\alpha_{kj}| + \beta) + \zeta \| \mathbf{D}^T \mathbf{D} - \mathbf{I}_K \|_F^2 + \eta \sum_{k=1}^K (\| \mathbf{D}_k \|_2^2 - 1)^2.
\]

The third term was introduced in a related work [21] to encourage low mutual coherence and Gram matrix norm of D, properties which are known to have a direct impact on the speed of sparse coding algorithms such as Iterative Shrinkage [8], and on the success of sparse coding formulations in recovering the correct sparse solutions [10], [23]. The last term in (6) is just a normalization one.

The SC step for the new sparse model (6) can be approximately solved using the Local Linear Approximation technique (LLA) [25], usually requiring less than 15 iterations to converge. Each iteration of LLA can be recast as a weighted instance of the standard SC step for (1), which can be solved efficiently using the tools already mentioned in Section II. The DU step is done using a Newton-like iteration similar to the one used in [12]. See [21] for details.

IV. EXPERIMENTAL RESULTS

In the following experiments the data X are 8×8 patches drawn from the 2600 the Pascal VOC2006 testing subset,\(^3\) converted to grayscale in the empirical one when using MOL, \(\lambda_{k=1}^K\), against the globally fitted Laplacian and MOL ones shown in Figure 1a, \(p_k\), and the ones fitted specifically to \(A_k^+, p_k^+\) and \(p_k^-\). The horizontal axis is sorted by increasing \(D\) for all \(p_k\), in all cases, and even \(p_k^+\) in most of the cases. This shows that MOL, with only two parameters, is a much better model than \(K\) Laplacians (requiring \(K\) parameters) fitted specifically to each atom. Whether these improvements have a practical impact is explored in the sequel.

B. Recovery of noisy sparse signals

Here we compare the active set recovery properties of the MOL prior, compared to those of the \(\ell_1\)-based one, on data for which the assumption \(|A_j| \leq L\) is made to hold exactly for all \(j\), for a small \(L\). To this end, we obtain sparse approximations to each patch \(X\), using the \(\ell_1\)-based Orthogonal Matching Pursuit algorithm (OMP) [19] on \(D\), and record the resulting active sets \(A_j\) as ground truth. The data is then contaminated with additive Gaussian noise of variance \(\sigma\) and the recovery is performed using the denoising formulation of BP, \(A_{\ell_1} = \arg \min_{\mathbf{A} \in \mathbb{R}^{M \times N}} \| \mathbf{A} \|_1\) s.t. \(\| \mathbf{X} - \mathbf{D}\mathbf{A} \|_2^2 \leq C\sigma^2\), and the “BP equivalent” for the MOL prior, \(A_{\text{MOL}} = \arg \min_{\mathbf{A} \in \mathbb{R}^{M \times N}} \sum_k \log (|\alpha_{kj}| + \beta)\) s.t. \(\| \mathbf{X} - \mathbf{D}\mathbf{A} \|_2^2 \leq C\sigma^2\). Here we use \(C = 1.32\), which is a standard value in denoising applications (e.g., [18]).

For each sample \(j\), we measure the error of each method in recovering its active set as the Hamming distance between the true and estimated support of its reconstruction coefficients. The accuracy of the method is then given as the percentage of the samples for which the error falls below a certain threshold \(T\), \(\text{PSNR}(L, T)\). Results are shown in Figure 1f for \(L = 5, T = 2\), for various values of \(\sigma\). Given the estimated active sets, the clean patches are estimated using least squares (which is the standard procedure for denoising when the active set is determined). We then measure the PSNR of the estimated patches with respect to the true ones. The results are shown as yellow lines in Figure 1e, again for various values of \(\sigma\). The red lines show the same results with a reduced mutual coherence (RMC) dictionary \(D\) learned using the additional terms included in (6). As can be observed, the MOL-based recovery is significantly better, specially in the high SNR case. It can also be observed that using the RMC dictionary consistently improves the results in all cases. Again, we refer the reader to [21] for details on this line of research.

C. Recovery of real signals with simulated noise

This experiment is an analogue of the previous one when the data are the original natural image patches. Since for this case the sparsity assumption is only approximate, and no ground

\(^3\)www.pascal.inrialpes.fr/challenges/VOC/databases.html#VOC2006

\(^4\)Similar results to those shown here are also obtained for other patch sizes. We choose not to include them due to space constraints.
truth is available for the active sets, we compare the different methods in terms of their denoising performance. In this case we include results for the $\ell_0$-based OMP algorithm as it is the one used to obtain state-of-the-art results in image denoising (see e.g. [1], [18]). The results in Figure 1g show that by using MOL we get the best of both the $\ell_1$ and $\ell_0$-based methods, by improving on OMP in low and high SNR regions while bridging the gap between $\ell_0$ and $\ell_1$ for mid-SNR. We also see that the use of a RMC dictionary (red lines) is not an advantage for this case, except for very high SNR. These results are preliminary, and further research is needed to assess whether imposing low mutual coherence can be advantageous for denoising tasks.

V. CONCLUDING REMARKS

A new prior for sparse modeling was introduced in this work, using tools from universal coding, whose significant theoretical and practical advantages over traditional regularization terms were shown. We also equipped our new model with an additional term which encourages incoherence in the learned dictionaries, a property which was also shown to improve the reconstruction properties of the model. The critical properties of the proposed model, such as increased stability of the active set and better sparse approximation, hint to the possible implications of this model for classification tasks such as those described in [17]. We are currently investigating this and results will be reported elsewhere.

ACKNOWLEDGEMENTS

Work partially supported by NGA, ONR, DARPA, ARO and FUNDACIBA-ANTEL. F.L. collaborated with this work while at the University of Minnesota.

REFERENCES

[23] J. A. Tropp. Greed is good: Algorithmic results for sparse approxima-

Fig. 1: (a) Empirical distribution of reconstruction coefficients $\alpha$ for image patches and best Laplacian and MOL fitting distributions. (b) Histogram of the $K$ different $\theta$ values obtained for each row $A^k$. (c) Empirical ($w^0_k$) and fitted Gamma ($w^1_k$) distributions of $\theta_k$ for some atoms $k$, when estimated from random subsamples of $X$. (d) Differences between the KLD for the best fitting distributions computed per atom. (e),(f) Reconstruction PSNR and active set recovery accuracy $E(L, T)$ of truly sparse signals for $L = 5$, $T = 2$. (g) Denoising of real data, using “normal” and RMC dictionaries. Results are relative to OMP using a normal dictionary. This figure is in colors.