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 choice 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 which has several theoretical and practical advantages over the more standard ℓ_0 or ℓ_1 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 (usually called dictionaries or codebooks), 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 ℓ_0 and ℓ_1-based techniques in practice. This new model is complemented by imposing incoherence in the learned dictionary.

II. Sparse modeling

Let $\mathbf{X} \in \mathbb{R}^{M \times N}$ be a set of N column data samples $\mathbf{X}_j \in \mathbb{R}^{M}$, $\mathbf{D} \in \mathbb{R}^{M \times K}$ be a dictionary of K atoms represented as columns $\mathbf{D}_k \in \mathbb{R}^{M}$, and $\mathbf{A} = \{\alpha_{kj}\} \in \mathbb{R}^{K \times N}, \mathbf{A}_j \in \mathbb{R}^{K}$, be a set of reconstruction coefficients such that $\mathbf{X} = \mathbf{DA}$. We also use $\mathbf{A}^k$ to denote the k-th row of $\mathbf{A}$, which corresponds to the coefficients associated to the k-th atom in $\mathbf{D}$. For each $j = 1, \ldots, N$ we define the *active set* of $\mathbf{A}_j$ as $\mathcal{A}_j = \{k : \alpha_{kj} \neq 0\}$, and $\|\mathbf{A}_j\|_0 = |\mathcal{A}_j|$ as its cardinality. 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,

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

where $\|\cdot\|_F$ denotes Frobenius norm. 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, under certain conditions on $\mathbf{D}$ and $\mathbf{X}$, the solutions to the ℓ_0 and ℓ_1-based sparse coding problems coincide [4].

Since the problem in (1) is non-convex in $(\mathbf{A}, \mathbf{D})$, the standard approach to find an approximate solution is alternate minimization. Starting with an initial dictionary $\mathbf{D}^{(0)}$, the following sequence of subproblems is repeated until convergence,

$$\begin{aligned} \text{SC} : \mathbf{A}^{(t+1)} &= \arg\min_{\mathbf{A}} f(\mathbf{X}, \mathbf{D}^{(t)}, \mathbf{A}) \\ \text{DU} : \mathbf{D}^{(t+1)} &= \arg\min_{\mathbf{D}} f(\mathbf{X}, \mathbf{D}, \mathbf{A}^{(t+1)}), \end{aligned}$$

where $f(\cdot)$ is the cost function in (1) and DU stands for Dictionary Update. The SC problem can be solved efficiently using for example Iterative Shrinkage [8] or LARS [9]. The DU step can be done using for example MOD [12].

III. Universal models for sparse coding

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

$$\mathbf{A}^* = \max_{\mathbf{A}} \log p(\mathbf{X}|\mathbf{A}) + \log p(\mathbf{A}), \tag{2}$$

where $p(\mathbf{X}|\mathbf{A}) \propto \exp(-\frac{1}{2\sigma^2}\|\mathbf{X} - \mathbf{DA}\|_2^2)$, and the prior on $\mathbf{A}$ is IID Laplacian with mean 0 and inverse-scale parameter θ, $p(\mathbf{A}) \propto \exp(-\theta \|\mathbf{A}\|_1)$. 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 $\hat{\lambda}$ (or $\hat{\theta}$, for given σ^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 θ can be different for each atom k and, furthermore, where each of these θ^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 with unknown parameter θ 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 $\mathbf{A}$ which can fit each $\mathbf{A}_j$ as well as the model in $\mathcal{M}$ that can be fitted to $\mathbf{A}_j$ 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 the subject of 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 $\alpha^n = (\alpha_1, \ldots, \alpha_n)$.

For given data α^n we measure the goodness of fit of a model $q(\cdot)$ with respect to $\mathcal{M}$ using the *codelength regret,*[1]

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

where $\hat{\theta}(\alpha^n)$ is the Maximum Likelihood Estimator of θ for the observed α^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, $q^*(\alpha^n) = \log p(\alpha^n|\hat{\theta}(\alpha^n))/\mathcal{C}(\mathcal{M}, n)$. The normalization constant $\mathcal{C}(\mathcal{M}, n) = \int_\Theta p(\alpha^n|\theta)d\theta$ 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 $\mathcal{C}(\mathcal{M}, n)$ for *any* sequence α^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

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

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

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

Since the regret in this case equals

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

where $D(\cdot||\cdot)$ is the Kullback-Leibler divergence (KLD), it is also trivial to see that $\mathcal{R}(q^w, q)$ 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 κ and β 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 θ. 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 θ (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

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

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

$$\hat{\kappa} = 2(\hat{\mu}_2 - \hat{\mu}_1^2)/(\hat{\mu}_2 - 2\hat{\mu}_1^2) \quad \text{and} \quad \hat{\beta} = (\hat{\kappa} - 1)\hat{\mu}_1, \qquad (5)$$

where $\hat{\mu}_j = \sum_{i=1}^n |\alpha_i|^j$ are the j-th non-central absolute sample moments. The role of β 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

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

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 ℓ_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 ℓ_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 ℓ_1-term at recovering sparse signals.[2] 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

[1] It is a standard assumption in universal coding to consider the codelengths given by the *Shannon code*, which assigns a codeword of length $-\log p(x)$ to a random value x with probability $p(x)$[7].

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

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 $\mathbf{D}$ into the following formulation

$$
\begin{aligned}
(\mathbf{A}^*, \mathbf{D}^*) \; = \; & \arg\min_{\mathbf{A},\mathbf{D}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \tau \sum_{j=1}^{N} \sum_{k=1}^{K} \log\left(|\alpha_{kj}| + \beta\right) \\
& + \zeta \left\|\mathbf{D}^T\mathbf{D} - \mathbf{I}_K\right\|_F^2 + \eta \sum_{k=1}^{K} (\|\mathbf{D}_k\|_2^2 - 1)^2. \quad (6)
\end{aligned}
$$

The third term was introduced in a related work [21] to encourage low mutual coherence and Gram matrix norm of $\mathbf{D}$, 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 $\mathbf{X}$ are $8{\times}8$ patches drawn from the 2600 the Pascal VOC2006 *testing* subset,[3] converted to grayscale in the $[0, 1]$ range.[4] We use a dictionary $\mathbf{D}$ with $K{=}256$ atoms trained to the VOC2006 *training* subset 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], [17].

A. MOL as a prior for reconstruction coefficients

We begin by evaluating the performance of the Laplacian and MOL models for fitting a single global distribution to the whole matrix $\mathbf{A}$. We compute $\mathbf{A}$ using the Basis Pursuit formulation (BP)[6] to obtain an exact reconstruction, $\min \|\mathbf{A}\|_1 \,\text{s.t.}\, \mathbf{X} = \mathbf{D}\mathbf{A}$, and then restrict our study to the nonzero elements of $\mathbf{A}$. Here $\mathbf{X}$ corresponds to all $8{\times}8$ patches from the 2600 *testing* VOC2006 images.

The empirical distribution of $\mathbf{A}$, p_{E}, is plotted in Figure 1a along with the best fitting Laplacian, p_{L}, and MOL, p_{M}, distributions. The MLE for the Laplacian fit is $\hat{\theta} = N_1/\left\|\mathbf{A}\right\|_1 = 27.2$ (here N_1 is the number of nonzero elements in $\mathbf{A}$). For MOL, using (5), we obtained $\kappa = 2.9$ and $\beta = 0.07$. The much better fitting observed in Figure 1a is further confirmed by a much smaller KLD of the fitted distribution with respect to the empirical one when using MOL, $D(p_{\mathrm{E}}||p_{\mathrm{M}}) = 0.04$, instead of a Laplacian, which yields $D(p_{\mathrm{E}}||p_{\mathrm{L}}) = 0.30$. As a reference, the empirical entropy of the data is $H(p_{\mathrm{E}}) = 3.00$ bits.

Figure 1b shows the histogram $h(\hat{\theta})$ of the $K = 256$ different values of $\hat{\theta}$ when fitted to each $\mathbf{A}^k$, $\{\hat{\theta}^k\}_{k=1}^K$. Figure 1c shows the empirical distribution of *each* $\hat{\theta}^k$ for some k, w_{E}^k, when computed from 20000 random subsamples of $\mathbf{X}$ of size 100, and corresponding best fitting Gamma distributions, w_{Γ}^k. Since the optimal $\hat{\theta}$ varies across samples, we expect the universal coding approach to perform well also on a per-atom basis. This is confirmed in Figure 1d, which shows the KLD between the empirical distribution of each $\mathbf{A}^k$, p_{E}^k, against the globally fitted Laplacian and MOL ones shown in Figure 1a, p_{L} and p_{M}, and the ones fitted specifically to $\mathbf{A}^k$, p_{L}^k and p_{M}^k. The horizontal axis is sorted by increasing $D(p_{\mathrm{E}}^k||p_{\mathrm{L}}^k) - D(p_{\mathrm{E}}^k||p_{\mathrm{M}})$. As can be seen, the KLD for the *global* p_{M} is significantly smaller than p_{L} in all cases, and even p_{L}^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 ℓ_1-based one, on data for which the assumption $|\mathcal{A}_j| \leq L$ is made to hold exactly for all j, for a small L. To this end, we obtain sparse approximations to each patch $\mathbf{X}_j$ using the ℓ_0-based Orthogonal Matching Pursuit algorithm (OMP) [19] on $\mathbf{D}$, and record the resulting active sets $\mathcal{A}_j$ as ground truth. The data is then contaminated with additive Gaussian noise of variance σ and the recovery is performed using the denoising formulation of BP, $\mathbf{A}^{\ell_1} = \arg\min_{\mathbf{A}} \|\mathbf{A}\|_1$, s.t. $\|\mathbf{X} - \mathbf{D}\mathbf{A}\|^2 \leq C\sigma^2$, and the "BP equivalent" for the MOL prior, $\mathbf{A}^{\mathrm{MOL}} = \arg\min_{\mathbf{A}} \sum \log\left(|\alpha_{kj}| + \beta\right)$ s.t. $\|\mathbf{X} - \mathbf{D}\mathbf{A}\|^2 \leq C\sigma^2$. 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, $E(L, T)$. Results are shown in Figure 1f for $L = 5, T = 2$, for various values of σ. 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 σ. The red lines show the same results with a reduced mutual coherence (RMC) dictionary $\mathbf{D}$ 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] http://pascallin.ecs.soton.ac.uk/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 constrains.

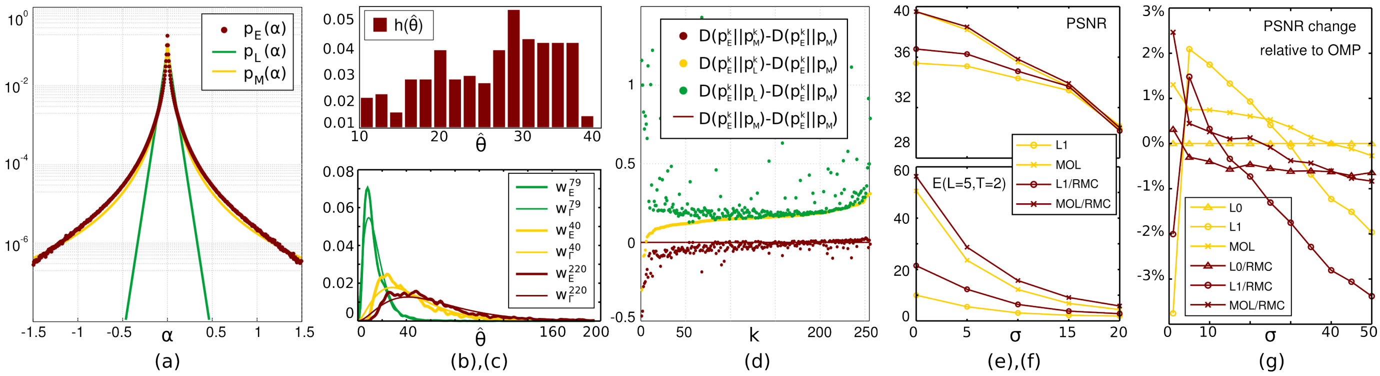

Fig. 1: (a) Empirical distribution of reconstruction coefficients α for image patches and best Laplacian and MOL fitting distributions. (b) Histogram of the K different $\hat{\theta}$ values obtained for each row $\mathbf{A}^k$. (c) Empirical (w_{E}^k) and fitted Gamma (w_{Γ}^k) distributions of $\hat{\theta}^k$ for some atoms k, when estimated from random subsamples of $\mathbf{X}$. (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.

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 ℓ_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 ℓ_1 and ℓ_0-based methods, by improving on OMP in low and high SNR regions while bridging the gap between ℓ_0 and ℓ_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

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

[7] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley and Sons, Inc., 2 edition, 2006.

[8] 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.

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

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

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

[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. *J. 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] 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.

[17] 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.

[18] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans. IP*, 17(1):53–69, Jan. 2008.

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

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

[21] I. Ramirez, F. Lecumberry, and G. Sapiro. Sparse modeling with universal priors and learned incoherent dictionaries. Submitted to NIPS, 2009.

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

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

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

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