Application of Concave/Schur-Concave Functions to the Learning of Overcomplete Dictionaries and Sparse Representations.

K. Kreutz-Delgado and B.D. Rao
Electrical and Computer Engineering Dept.
University of California, San Diego
La Jolla, CA 92093-0407
{kkreutzd, brao}@ucsd.edu

Abstract

Given a very overcomplete $m \times n$ dictionary of representation vectors a_i, $A = [a_1, ..., a_n]$, $n >> m$, an environmentally generated signal, y, can be succinctly represented within the dictionary by obtaining a sparse solution, x, to the linear inverse problem $Ax \approx y$ using various recently proposed methodologies. In particular, sparse solutions can be found by an appropriately regularized minimization of the error $e = y - Ax$. In this paper we briefly discuss our recent investigations into the use of concave/schur-concave functions as regularizing sparsity measures, and their application to the problem of obtaining sparse representations, x, of environmentally generated signals y, and the problem of learning evironmentally adapted overcomplete dictionaries.

1 Introduction

Background. In recent years, there has been a significant expansion in the understanding of the role of signal features for representing both entire signal spaces *and* specific signal instantiations. In the first instance it is desirable to have a set of feature/signal vectors that can well represent the totality of signals provided by the environment of interest (the "signal soup").

Environmentally generated signals typically have significant statistical structure, and can be represented by a set of basis vectors spanning a lower dimensional submanifold of meaningful signals [8, 32]. These environmentally-meaningful representation vectors can be obtained by maximizing the mutual information between the set of these vectors (the dictionary) and the signals generated by the environment [4, 1, 5, 22, 36, 33]. This procedure can be viewed as a natural generalization of Independent Component Analysis (ICA) [4, 5]. As initially developed, this procedure usually results in obtaining a *minimal* spanning set of spanning vectors (i.e., a true basis). More recently, the desirability of obtaining "overcomplete" sets of vectors (or "dictionaries") has been noted [22, 15, 3, 16, 6, 28]. For example, projecting measured noisy signals onto the signal submanifold spanned by a set of dictionary vectors results in noise reduction and data compression [6, 7]. These dictionaries can be structured as a set of true bases from which a single basis is to be selected to represent the measure signal(s) of interest [3], or as a large set of individual vectors [16], perhaps even unorganized in any particular way [22, 15, 28].

The problem of determining a representation from an overcomplete dictionary, $A = [a_1, ..., a_n]$, $n >> m$, for a specific signal measurement, y, is equivalent to solving a underdetermined inverse problem, $Ax \approx y$. The standard least squares solution to this problem has the (at times) undesirable feature of involving *all* the dictionary vectors in the solution[1] (the "spurious artifact" problem), and does not allow for the extraction of a categorically or physically meaningful solution. That is, it is not generally the case that a least-squares solution yields a concise representation allowing for a precise semantic meaning[2].

If the dictionary is large and rich enough in representational power, a measured signal can be matched to a very few (perhaps even just one) dictionary words. In this manner we can obtain concise semantic content about objects or situations encountered in natural environments [8]. Thus, there has been a significant interest in finding "sparse" solutions, x, (solu-

[1] Later, we'll see that this fact comes as no surprise when the solution is interpreted within a Bayesian framework.

[2] Taking "semantic" here to mean categorically or physically interpretable.

0-7803-5148-7/98/$10.00©1998 IEEE

tions having a minimum number of nonzero elements) to the signal representation problem. Interestingly, matching a *specific* signal to a sparse set of dictionary words/vectors can be related to entropy *minimization* as a means of elucidating statistical structure [34]. Finding a sparse representation (based on the use of a "few" code/dictionary words) can also be viewed as a generalization of vector quantization where a match to a single "code vector" (word) is always sought (taking "code book" = "dictionary"). Indeed, we can refer to a sparse solution, x, as a sparse coding of the signal instantiation, y.

Outline of the Paper. In this paper, we briefly discuss our recent efforts to develop a analytic framework and algorithms for learning environmentally adapted overcomplete dictionaries and obtaining sparse signal representations given an overcomplete dictionary. Our approach is readily interpretable within a Bayesian framework, and is based on regularized minimization of the representation error $\|e\|^2$, $e = y - Ax$, where the regularizing functions are sparsity-enforcing "entropy-like" diversity measures that have the mathematical properties of being concave and Schur-concave. We also examine the problem of dictionary learning and the relationship to independent component analysis (ICA).

We present an efficient algorithm for searching through a large dictionary to find sparse codes [12]–[13]. Unlike forward sequential search algorithms which search serially through the dictionary to obtain a sparse coding [16], this algorithm searches parallely through the dictionary to find an entropy minimizing sparse solution. The algorithm can be viewed as a generalization of 1-norm optimization-based methods for finding sparse codes [15, 36, 2]. Specifically, using the framework of majorization and concave function theory [17, 27] we have developed a class of parameterized entropy-like measures and associated affine scaling transformation (AST)-based diversity (entropy-like) functional minimization algorithms that yield sparse solutions to the signal representation problem.

2 Bayesian Framework

A Bayesian interpretation is obtained from the generative signal model,

$$y = Ax + \nu\,, \tag{1}$$

where x has the parameterized pdf,

$$P_p(x) = Z_p^{-1} e^{-d_p(x)}\,, \quad Z_p = \int e^{-d_p(x)} dx\,, \tag{2}$$

with parameter vector, p, and and the noise ν is assumed to be normally distributed, $P_\nu \sim \mathrm{N}(0, \frac{\lambda}{2} \cdot I)$. We treat A and p as deterministic, but possibly unknown parameters, and thus x and y are jointly distributed as $P(x, y; p, A)$. We also assume that $P(x; p, A) = P(x; p) = P_p(x)$. Bayes' rule yields,

$$P(x|y; p, A) = \frac{1}{\beta} P(y|x; p, A) \cdot P(x; p, A) = \frac{1}{\beta} P_\nu(y - Ax) \cdot P_p(x)\,,$$

where

$$\beta = P(y; p, A) = \int P(y|x; A) \cdot P(x; p) dx\,.$$

We will refer to x henceforth as the *source vector* or as the *representation vector* of the measured signal y.

Prespecified Dictionary. When the prior $P_p(s)$ and the dictionary A are both prespecified (the known dictionary case), maximizing $P(x|y; p, A)$ (equivalently, minimizing its negative logarithm) leads to the optimization problem,

$$\hat{x} = \arg\min_x \left\{ \|Ax - y\|^2 + \lambda\, d(x) \right\}\,, \tag{3}$$

In the low noise limit, $\lambda \to 0$, this becomes

$$\hat{x} = \min_{Ax=y} d(x)\,. \tag{4}$$

That sparse solutions arise from an appropriate choice of the regularizing function $d(x)$ has been noted by a variety of researchers [25, 26, 21, 22, 36, 15, 29].

We see, then, that within the Bayesian framework such a solution corresponds to a *maximum a posteriori* (MAP) estimate of the representation x given a dictionary, A, a prior on the space of representations, $P_p(x)$, and noise variance $\lambda/2$. A prespecified dictionary can be determined in a variety of ways, e.g., as a wavelet frame, an overcomplete local cosine dictionary, a previously learned dictionary, or (even) from a polling of human experts.

Unknown Dictionary. Let us assume that $P_p(x)$ defines a rich class of parameterized priors which can well model the true (but unknown) representation prior of x. Lack of knowledge of the prior then corresponds to not knowing a correct value of the parameter vector p. For *unknown* p and/or A, it is required to learn values of p and the dictionary A "best adapted" to the statistics of the environment generating the observed signal y. In essence, we want to find a generative model of the form (1) that best explains the observations. Note that the source vector x is unknown (the "blind source problem") which makes the task of estimating p and A somewhat problematic. Assuming that we

can collect a sufficiently representative (and large) sample set of independent observations $Y^N = \{y_1, \cdots y_N\}$ generated by the respective sequence of independent source vectors $X^N = \{x_1, \cdots, x_N\}$, then

$$P(X^N|Y^N; p, A) = \prod_{\ell=1}^{N} P(x_\ell|y_\ell; p, A)\,.$$

We can obtain *maximum likelihood* estimates of p, A, and X^N given Y^n by solving the problem

$$\min_{p,A,X^N} -\log p(X^N|Y^N; p, A) = \min_{p,A,x^N} -\sum_{\ell=1}^{N} \log p(x_\ell|y_\ell; p, A)\,.$$

This can be written in terms of the sample average, $< \cdot >= \frac{1}{N}\sum_\ell(\cdot)$, as

$$\min_{p,A,x^N} \langle -\log p(x|y; p, A)\rangle\ ,$$

which with the generative model (1) and fixed p gives the optimization problem

$$\widehat{A} = \arg \min_{A,x_1,x_2,\cdots} \langle \{\|Ax - y\|^2 + \lambda\, d(x)\}\rangle\ , \qquad (5)$$

where $\langle\cdot\rangle$ indicates an averaging over environmental samples $y \in \{y_1, y_2, \cdots\}$ generated by the "source vectors" $\{x_1, x_2, \cdots\}$. This procedure results in our *learning* a dictionary adapted to the environment [36, 15]. It is evident that choice of $d(x)$ affects the nature of the learned dictionary A and, given A, any particular sparse solution to (5). This approach is analyzed for the case of the ℓ_1-norm prior in [21, 22, 15, 20].

This procedure can also be given interesting information theoretic interpretations, and conditions can be given that ensure optimality of the resulting learned probability distributions in terms of the Kullback-Liebler distance to the true underlying probabilities [19, 1, 23, 24, 31]. Further, the optimization problem (5) is equivalent to maximizing the mutual information between x and the observation y [1, 5].

Ideally, optimizing over the parameterization p will provide a good parametric fit of $P_p(x)$ to the true underlying prior probability density function describing the source vectors. Thus, we see the potential desirability of having a relatively large family of distributions $P_p(x)$. The problem of selecting 'an 'optimal" choice of p is known as the problem of hyperparameter selection in the Bayesian estimation literature [35, 36].

3 Concave Regularizers.

Recently we have argued that appropriate regularizing functions for obtaining sparse solutions to problems (3) or (4) should be separable, concave, and Schur-concave. In the remainder of the paper we will discuss the implications of this restriction for algorithm development and as interpreted within the Bayesian framework.

Separability and ICA. A detailed discussion of Schur-concavity, concavity, and separability, and the utility of these properties for obtaining sparse solutions can be found in [11, 14]. These references also discuss several classes of regularizing functions having these properties.

Separable functions are particularly useful for constructing regularization functions that are Schur-concave. They obey the property that

$$d(x) = \sum_{i=1}^{n} \phi(x[i])\,,$$

where $x[i]$ is the i^{th} component of $x \in \mathsf{R}^n$. Note that separability of $d(x)$ corresponds to *factorizability* of $P_p(x)$,

$$P_p(x) = P_p(x[1]) \cdots P_p(x[n])\,.$$

- Thus *separability* of $d(x)$ corresponds to the assumption of *independent components* of x.

We see that from a Bayesian perspective, separability of $d(x)$ corresponds to a generative model for y that *assumes a source, x, with independent components.* With this assumption, we are working within the framework of Independent Component Analysis (ICA) [19, 1, 23, 24, 31] and an algorithm that optimizes (5) assuming the validity of the generative model (1) can be viewed as an ICA algorithm.

Note that relaxing the restriction of separability generalizes the generative model to the case where the source vector, x, has *dependent components.* We can reasonably call an approach based on a non-separable diversity measure $d(x)$ a *Dependent Component Analysis* (DCA). Preliminary results appear to show that this relaxation significantly complicates the analysis and development of optimization algorithms.

Concave Regularizers and Sparse Solutions: Bayesian Interpretation. Perhaps the paradigmatic separable concave/Schur-concave regularizing functions are given by the class of $\ell_{p\le 1}$ sparsity measures [30, 11],

$$d_p(x) = \mathrm{sgn}(p)\sum_{i=1}^{n} |x[i]|^p,\ \text{scalar } p \le 1\,. \qquad (6)$$

The priors (2) induced by this choice are *supergaussian*, (i.e., have postive kurtosis). Such priors are known to favor solutions, x, with zero entries [18].

- More generally, supergaussian priors will be induced by the class of concave/Schur-concave regularizing functions.

This is a fact consistent with our earlier arguments that such functions enforce sparse solutions [30, 11]. Note that $d_p(x)$ for $p = 2$ (which is *not* in the class (6)) corresponds to a gaussian prior and would result in the standard 2-norm regularizing function $d(x) = \|x\|^2$.

It is well-known that the gaussian prior is a maximum entropy distribution [9] (subject to constraints on the first two moments) and corresponds to a maximally non-committal distribution of probabilities over the entries of x. From this perspective, it is not surprising that the 2-norm regularized optimization (3) generally yields nonsparse solutions. That the $p \leq 1$ (including p negative) supergaussian case yields sparse solutions has been noted by many researchers. In particular, the Laplacian prior given by $p = 1$ has garnered much interest [2, 15]. Plots of $d_p(x)$ for $p = 0.5, 1.0, 2.0$ are given in [18].

Note that from the perspective of ICA, it is *separability* that allows us to consider (3) as yielding an ICA solution, while the *additional assumption of concavity/Schur-concavity* of $d(x)$ (resulting in supergaussianess of the prior $P(x)$) enforces sparsity.

4 Algorithms for Sparse Solutions.

Factorization-Based Problem Restatement. In our recent work we have proposed nonlinear algorithms for obtaining sparse solutions given a dictionary which correspond to solving problem (4), and more recently problem (3), using an appropriate separable, concave/Schur-concave prior [29, 30, 11]. Note that (3) has a regularizing term, $\lambda d(x)$, that in general is very difficult to handle (not being the standard mathematically tractable quadratic regularizer). However, our methodology has the following important interpretation.

- A solution to (3) can be found from solving a *sequence of mathematically tractable 2-norm optimizations.*

Specifically, the sequence of optimizations is of the form,

$$q_{k+1} = \arg\min_q \|AW_{k+1}q - b\|^2 + \lambda\|q\|^2 \qquad (7)$$

with $x_k = W_k q_k \rightarrow \hat{x}$, as $k \rightarrow \infty$ where $\hat{x}$ is a local solution of (3). The key to implementing this algorithm is based on a certain *gradient factorization* of the regularizing function $d(x_k)$, which yields the weighting matrix $W_{k+1} = W(x_k)$. Further discussion of this algorithm is given in [30, 29, 11]. Simulations showing the behavior of the algorithm in the low noise case can be found in [30].

Dictionary Learning. A very interesting problem is to learn an environmentally adapted dictionary, A, given a statistically representative sample of observations, Y^N, drawn from the environment. This problem has been recognized and posed for the Laplacian prior. This is a difficult optimization problem often solved using markovian sampling techniques (such as simulated annealing and other Markov Chain Monte Carlo (MCMC) algorithms). It has been noted that learning dictionaries for "natural images" in this manner using the Laplacian prior results in Gabor function-like basis vectors [21, 22, 15], and recently it has been argued (from the perspective that solving the inverse problem (3) is equivalent to maximizing mutual information) that this is to be expected on the grounds that Gabor functions are locally optimal in an information theoretic sense [20].

We are beginning to explore the development of learning algorithms for the class of separable, concave/Schur-concave regularizing functions and hope to exploit the simplifications afforded by the gradient factorization-based algorithm used in the known-dictionary case. Indeed, our preliminary investigations suggests that this will proved to be a fruitful line of inquiry, and we hope to report results in the near future.

5 Conclusions

We have discussed the utility of a Bayesian framework for understanding the use of separable, concave/Schur-concave regularizing functions in solving linear inverse problems for sparse solutions. In particular, separability corresponds to an independent source component assumption, while concavity/Schur-concavity corresponds to the use of a supergaussian, sparsity enforcing prior. We have discussed the fixed dictionary case and the obtaining of sparse solutions as the solution of the regularized minimization problem (3) and noted that we can transform this seemingly intractable regularized optimization problem into a sequence of solvable 2-norm regularized optimization problems, thereby obtaining a sparse, locally optimal solution. We also noted that some recent successful

results on learning an environmentally adapted dictionary have been reported in the literature, and in particular when applied to natural images have resulted in Gabor function-like basis vectors that are arguable optimal in an information theoretic sense. We are also currently working on the problem of dictionary learning, in particular we are interested in the question of whether the gradient factorization-based algorithms we have recently developed in the fixed-dictionary case can be extended to accomplish environmentally-adapted dictionary learning and are actively pursuing this line of investigation.

References

[1] A. Bell & T. Sejnowski, "An Information Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Comp.*, 7:1129-59, 1995.

[2] S. Chen, D. Donoho, & M. Saunders, "Atomic Decomposition by Basis Pursuit," Technical Report, Department of Statistics, Stanford University, 1996.

[3] R. Coifman & M. Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection," *IEEE Trans. Inf. Theory*, IT-38(2):713-18, 1992.

[4] P. Comon, "Independent Component Analysis: A New Concept?" *Signal Proc.*, Vol. 36, pp. 287-314, 1994.

[5] G. Deco & D. Obradovic, *An Information-Theoretic Approach to Neural Computing*, Springer 1996.

[6] D. Donoho, "On Minimum Entropy Segmentation," in *Wavelets: Theory, Algorithms, and Applications*, C. Chui et al., Ed.'s, Academic Press, 1994.

[7] D. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, May 1995, vol.41, (no.3):613-27.

[8] D. Field, "What is the Goal of Sensory Coding," *Neural Computation*, Vol. 6, pp. 559-99, 1994.

[9] E.T. Jaynes, "Prior Probabilities," *IEEE Trans. Sys. Man Cybernetics*, SSC-4(3):227-241, Sept. 1968.

[10] R. Kamimura, "Information Controller to Maximize and Minimize Information," *Neural Computation*, 9(6):1357-80, 1997.

[11] K. Kreutz-Delgado & B.D. Rao, "A General Approach to Sparse Basis Selection: Majorization, Concavity, and Affine Scaling," UCSD CIE Report 1997. Submitted to *Signal Processing.*

[12] K. Kreutz-Delgado & B.D. Rao, "Measures and Algorithms for Best Basis Selection," *1998 ICASSP.*

[13] K. Kreutz-Delgado & B.D. Rao, "Gradient Factorization Based Algorithm for Best-Basis Selection," *The 8th IEEE Signal Processing Workshop*, Bryce Canyon, UT, 1998.

[14] K. Kreutz-Delgado & B.D. Rao, "Sparse Basis Selection, ICA, and Majorization: Towards a Unified Perspective," submitted to *ICASSP-99.*

[15] M. Lewicki & T. Sejnowski, "Learning Nonlinear Overcomplete Representations for Efficient Coding," February 1998, Preprint. Submitted to *Neural Computation.*

[16] S. Mallat & Z. Zhang, "Matching Pursuits with Time-Frequency Dictionaries," *Trans. IEEE ASSP*, 41(12):3397-416, 1993.

[17] A. Marshall & I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, 1979.

[18] M.N. Murthi & B.D. Rao, "Towards a Synergistic Multistage Speech Coder," *ICASSP-98.*

[19] J.-P. Nadal and N. Parga, "Nonlinear Neurons in the Low Noise Limit: a Factorial Code Maximizes Information Transfer," *Network*, 5(4):565-81, November 1994.

[20] K. Okajima, "The Gabor Function Extracts the Maximum Information from Input Local Signals," *Neural Networks*, 11(3):435-39, April 1998.

[21] B.A. Olshausen and D. Field, "Learning Efficient Linear Codes for Natural Images: The Roles of Spareness, Overcompleteness, and Statistical Independence", *SPIE Proc.: Human Vision and Electronic Imaging*, 2657:132-8, 1996.

[22] B. Olshausen & D. Field, "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?," December 1996, Preprint.

[23] B.A. Pearlmutter and L.C. Parra, "Maximum likelihood blind source separation: a context-sensitive generalization of ICA," *Proc. NIPS-96*, p. 613-619, 1996).

[24] D.T. Pham, "Blind Separation of Instantaneous Mixture of Sources via an Independent Component Analysis," *IEEE Trans. Signal Processing*, 44(11):2768-79, 1997.

[25] R.C. Puetter, "Pixons and Bayesian Image Reconstruction". *Proc. SPIE: Image Reconstruction and Restoration*, 2302:112–31, 1994.

[26] R.C. Puetter, "Information, Language, and Pixon-Based Image Reconstruction", *Proc. SPIE: Digital Image Recovery and Synthesis III*, 2827:12–31, 1996.

[27] R. Rockafellar, *Convex Analysis*, Princeton, 1970.

[28] B.D. Rao & K. Kreutz-Delgado, "Deriving Algorithms for Computing Sparse Solutions to Linear Inverse Problems," *Proc. 1997 Asilomar Conference on Circuits, Systems, and Computers.*

[29] B.D. Rao & K. Kreutz-Delgado, "Basis Selection in the Presence of Noise," 1998 Asilomar Conference.

[30] B.D. Rao & K. Kreutz-Delgado, "An Affine Scaling Methodology for Best Basis Selection," *IEEE Trans. Signal Processing*, January 1999.

[31] S.J. Roberts, "Independent Component Analysis: Source Assessment and Separation, a Bayesian Approach," *IEE Proc.-Vis. Image Signal Process.* 145(3):149-53, 1998.

[32] D. Ruderman, "The Statistics of Natural Images," *Network: Computation in Neural Systems*, Vol. 5, pp. 517-48, 1994.

[33] K. Wang, C. Lee, & B. Juang, "Selective Feature Extraction via Signal Decomposition," *IEEE Signal Processing Letters*, 4(1):8-11, 1997.

[34] S. Watanabe, "Pattern Recognition as a Quest for Minimum Entropy," *Pattern Recognitio*, 13(5):381-87, 1981.

[35] Z. Zhou, R.M. Leahy, and J. Qi, "Approximate Maximum Likelihood Hyperparameter Estimation for Gibbs Priors," *IEEE Trans. Signal Proc.*, 6(6):844-61, 1997.

[36] S. Zhu, Y. Wu, & D. Mumford, "Minimax Entropy Principle and its Application to Texture Modeling," *Neural Comp.*, 9:1627-60, 1997.