Novel Algorithms for Learning Overcomplete Dictionaries*

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

Abstract

Using a Bayesian framework based on an assumption of a Convex/Schur-Convex (CSC) log-prior, together with an associated Affine-Scaling Transformation (AST) optimization algorithm, a signal vector, y, can be succinctly represented within a very overcomplete $m \times n$ dictionary of representation vectors a_i, $A = [a_1, ..., a_n]$, $n >> m$, dictionary by obtaining a sparse solution, x, to the linear inverse problem $Ax \approx y$. In this paper we outline how novel Approximate Maximum Likelihood (AML) and Maximum A Posteriori (MAP) overcomplete dictionary learning algorithms can be developed within the CSC/AST framework.

1 Introduction

It is well known that the stochastic generative model

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

can be used to develop algorithms enabling coding of $y \in \mathsf{R}^m$ via solving the inverse problem for a spare solution $x \in \mathsf{R}^n$ for the undercomplete ($n < m$) and complete ($n = m$) cases [1]. In recent years there has been a great deal of interest in obtaining sparse codings of y via this procedure for the *overcomplete* ($n > m$) case [12, 3]. In our earlier work we have shown that given an overcomplete dictionary, A, (with the columns of A comprising the dictionary vectors) a MAP estimate of the source vector, x, will yield a sparse coding of y in the low-noise limit if the negative log-prior, $-\log(P(x))$, is Concave/Schur-Concave (CSC)[18, 8]. For $P(x)$ factorizable into a product of marginal probabilities, the resulting code is also known to provide an Independent Component Analysis (ICA) representation of y. More generally, a CSC prior results in a sparse representation even in the non-factorizable case (with x then forming a "Dependent Component Analysis," or DCA, representation) [5, 9].

As noted by [13, 11], given iid data, $Y = Y^N = (y_1, ..., y_N)$, which is assumed to be generated by the model (1), a maximum likelihood estimate, $\widehat{A}_{\mathrm{ML}}$, of the unknown (but nonrandom) dictionary A can be determined as

$$\widehat{A}_{\mathrm{ML}} = \arg\max_A P(Y; A)\,.$$

This requires integrating out the unobservable iid source vectors, $X = X^N = (x_1, ..., x_N)$, in order to compute $P(Y; A)$ from the (assumed) known probabilities $P(x)$ and $P(\nu)$. In essence X is formally treated as a set of nuisance parameters which, in principle, can be removed via integration. However, because the prior $P(x)$ is generally taken to be supergaussian, this integration is intractable or computationally unreasonable. Thus approximations to this integration are performed which results in an approximation to $P(Y; A)$ which is then maximized with respect to Y. A new, better, approximation to the integration can then be made and this process is iterated until the estimate of the dictionary A has (hopefully) converged [13]. We refer to the resulting estimate as an Approximate Maximum Likelihood (AML) estimate of the dictionary A (denoted here by $\widehat{A}_{\mathrm{AML}}$). No formal proof of the convergence of this algorithm to the true maximum likelihood estimate, A_{ml}, has been given in the prior literature, but it appears to perform well in various test cases [13].

In this paper we examine the problem of dictionary learning within the framework of our recently developed CSC log-prior model-based sparse source vector learning approach which for a *known* overcomplete dictionary can be used to obtain sparse codes, both for the ICA (factorial code) and DCA (nonfactorial code)

*Research partially supported by NSF grant no. CCR-9902961

†Current address: ECE Dept., Høgskolen i Stavanger, Stavanger Norway.

0-7803-5700-0/99/$10.00©1999 IEEE

cases [18, 5, 6, 7, 17, 8]. Such sparse codes are found using FOCUSS, an affine scaling transformation (AST)-like interative algorithm which finds a sparse locally optimal MAP estimate of the source vector x for an observation y. Using these results, we can develop dictionary learning algorithms, both within the Approximate Maximum Likelihood framwork mentioned above and for obtaining a MAP-like estimate, $\widehat{A}_{\text{MAP}}$, of the (now assumed random) dictionary, A, assuming in the later case that the dictionary belongs to a compact submanifold corresponding to unit Frobenius norm. Under certain conditions, convergence to a local minimum of a MAP-loss function which measures a combination of functions of the discrepancy $e = (y - Ax)$ and the degree of sparsity in x can be rigorously proved.

2 Bayesian Dictionary Learning

Known Dictionary Model. A Bayesian interpretation is obtained from the generative signal model (1) by assuming that x has the parameterized (generally nongaussian) pdf,

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

with parameter vector p. Similarly, the noise ν is assumed to have a parameterized (possibly nonguassian) density $P_q(\nu)$ of the same form as (2) with parameter vector q. It is assume that x and ν have zero means and that their densities obey the property $d(x) = d(|x|)$, for $|\cdot|$ defined component-wise. This is equivalent to assuming that the densities are symmetric with respect to sign changes in the components of x, $x[i] \leftarrow -x[i]$, and therefore that the skews of these densities are zero. We also assume that $d(0) = 0$. With a slight abuse of notation, we allow the differing subscripts q and p to indicate that d_q and d_p may be *functionally* different as well as parametrically different. We refer to densities like (2), for suitable additional constraints on $d_p(x)$, as Hypergeneralized Gaussian Distributions [8, 9].

If we treat A, p, and q as *known* parameters, then x and y are jointly distributed as $P(x, y) = P(x, y; p, q, A)$. Bayes' rule yields,

$$P(x|y; p, A) = \frac{1}{\beta} P(y|x; p, A) \cdot P(x; p, A) = \frac{1}{\beta} P_q(y - Ax) \cdot P_p(x) \quad (3)$$

$$\beta = P(y) = P(y; p, q, A) = \int P(y|x) \cdot P_p(x) dx. \quad (4)$$

Usually the dependence on p and q is notationally suppressed and we write $\beta = P(y; A)$, etc. Given an observation, y, maximizing (3) with respect to x yields the MAP estimate $\widehat{x}$. This ideally results in a sparse coding of the observation, a requirement which places functional constraints on the probability density functions. Note that β is independent of x and can be ignored when optimizing (3) with respect to the unknown source vector x.

The MAP estimate equivalently is obtained from minimizing the the negative logarithm of $P(x|y)$, which is,

$$\widehat{x} = \arg\min_x d_q(y - Ax) + \lambda d_p(x), \quad (5)$$

where $\lambda = \gamma_p / \gamma_q$, and $d_q(y - Ax) = d_q(Ax - y)$ by our assumption of symmetry. The quantity $\frac{1}{\lambda}$ is interpretable as a signal-to-noise ratio (SNR). Furthermore, assuming that d_q and d_p are CSC, the term $d_q(y - Ax)$ in (5) encourages sparse residuals, $e = y - A\widehat{x}$, while the term $d_p(x)$ encourages sparse source-vector estimates, $\widehat{x}$. A given value of λ then determines a trade-off between residual and source-vector sparseness.

Note that $\lambda \to 0$ as $\gamma_p \to 0$ which (consistant with the generative model (1)) we refer to as the *low noise limit*. Because the mapping A is assumed to be onto, in the low noise limit the optimization (5) is equivalent to the linearly constrained problem,

$$\widehat{x} = \arg\min d_p(x) \quad \text{subject to} \quad Ax = y. \quad (6)$$

In the low-noise limit, no sparseness constraint is placed on the residuals $e = y - A\widehat{x}$. It is evident that the structure of $d_p(\cdot)$ is critical for obtaining a sparse coding, $\widehat{x}$, of the observation y, as extensively discussed in [5, 17].

In this paper, $d_p(x)$ is assumed to be CSD (enforcing sparse solutions to the inverse problem (1)) while ν is assumed to be Gaussian ($q = 2$).

Unknown, Deterministic Dictionary. The Maximum Likelihood Estimation framework treats the parameters as unknown but deterministic. Here, we now take the dictionary, A, to be the set of unknown parameters to be estimated from the observation set $Y = Y^N$. In particular, given Y^N the maximum likelihood estimate $\widehat{A}_{\text{ML}}$ is found from maximizing the likelihood function $L(A|Y^N) = P(Y^N; A)$. Under the assumption that the observations are iid, this corresponds to the optimization,

$$\widehat{A}_{\text{ML}} = \arg\max_A \prod_{k=1}^{N} P(y_k; A), \quad (7)$$

$$\begin{aligned} P(y_k; A) &= \int P(y_k, x; A) dx = \int P(y_k|x; A) \cdot P_p(x) dx \\ &= \int P_q(y_k - Ax) \cdot P_p(x) dx. \end{aligned} \quad (8)$$

Defining the sample average of a function $f(y)$ over the sample set $Y^N = (y_1, \cdots, y_N)$ by

$$\langle f(y) \rangle_N = \frac{1}{N} \sum_{k=1}^{N} f(y_k) ,$$

the optimization (7) can be equivalently written as

$$\widehat{A}_{\text{ML}} = \arg\min_A - \langle \log(P(y; A)) \rangle_N . \qquad (9)$$

Note that $P(y_k; A)$ is equal to the normalization factor β encountered earlier above, but now with the dependence of β on A and the particular sample, y_k, made explicit. The integration in (8) in general is intractible, and various approximations have been proposed to obtain an Approximate Maximum Likelihood estimate, $\widehat{A}_{\text{AML}}$ [13, 11].

Reference [13] proposes the approximation

$$P_p(x) \approx \delta(x - \widehat{x}_k(\widehat{A})) , \qquad (10)$$

where

$$\widehat{x}_k(\widehat{A}) = \arg\max_x P(y_k, x; \widehat{A}) , \qquad (11)$$

for $k = 1, \cdots, N$, assuming a current estimate, $\widehat{A}$, for A. With this approximation, the optimization (9) becomes,

$$\widehat{A}_{\text{AML}} = \arg\min_A \langle d_q(y - \widehat{A}\widehat{x}) + \lambda d_p(\widehat{x}) \rangle_N , \qquad (12)$$

which is an optimization over the sample average of the functional (5) encountered earlier. Updating our estimate for the dictionary,

$$\widehat{A} \leftarrow \widehat{A}_{\text{AML}} , \qquad (13)$$

we can iterate the procedure (11)–(12) until $\widehat{A}_{\text{AML}}$ has converged, hopefully (at least in the limit of large N) to $\widehat{A}_{\text{ML}} = \widehat{A}_{\text{ML}}(Y^N)$ as the maximum likelihood estimate $\widehat{A}_{\text{ML}}(Y^N)$ has well-known desirable asymptotic properties in the limit $N \to \infty$.

Performing the optimization in (12) for the $q = 2$ iid gaussian measurement noise case (ν gaussian with covariance $\frac{1}{\sigma^2} \cdot I$),

$$d_q(y - \widehat{A}\widehat{x}) = \frac{1}{2\sigma^2} \|y - \widehat{A}\widehat{x}\|^2 , \qquad (14)$$

we readily obtain the unique 'batch' solution,

$$\widehat{A}_{\text{AML}} = \Sigma_{y\hat{x}}^T \Sigma_{\hat{x}\hat{x}}^{-1} , \qquad (15)$$

$$\Sigma_{y\hat{x}} = \frac{1}{N} \sum_{k=1}^{N} y_k \widehat{x}_k^T , \quad \Sigma_{\hat{x}\hat{x}} = \frac{1}{N} \sum_{k=1}^{N} \widehat{x}_k \widehat{x}_k^T . \qquad (16)$$

This solution can be compared to the maximum likelihood estimate of A for the ideal case of *known* source vectors X,

$$\text{Known Source Vector Case:} \quad A_{\text{ML}} = \Sigma_{yx}^T \Sigma_{xx}^{-1} ,$$

which is, of course, not computable since the source vectors $X = (x_1, \cdots, x_N)$ are assumed to be 'hidden.'

Instead of using the explicit solution (15), which requires an often prohibitive $n \times n$ inversion, we can obtain A_{AML} iteratively via gradient descent on (12)/(14),

$$\begin{aligned} \widehat{A}_{\text{AML}} &\leftarrow \widehat{A}_{\text{AML}} - \alpha \Sigma_{k=1}^{N} e_k \widehat{x}_k^T , \qquad (17) \\ e_k &= \widehat{A}\widehat{x}_k - y_k , \quad k = 1, \cdots, N , \end{aligned}$$

for an appropriate choice of the (possibly adaptive) positive step-size parameter α. Note the distinction in (17) between $\widehat{A}$ and $\widehat{A}_{\text{AML}}$, the later hopefully converging to the batch estimate (15). The iteration (17) can be initialized as $\widehat{A}_{\text{AML}} = \widehat{A}$.

A general iterative dictionary learning procedure is obtained by nesting the interation (17) entirely within the iteration defined by repeatedly solving (11) every time a new estimate, $\widehat{A}_{\text{AML}}$, of the dictionary becomes available. As discussed in [13, 11], performing the optimization required in (11) is nontrivial.

References [16, 2] develop an effective algorithm for performing the optimization required in (11) for the case when ν is gaussian. This algorithm solves (11) using the Affine-Scaling Transformation (AST)-like algorithms recently proposed for the low noise case in [15, 5, 17] and extended via regularization to the nontrivial noise case in [16, 2]. For a current dictionary estimate, $\widehat{A}$, a solution to the optimization problem is provided by repeated iteration of the form,

$$x_k \leftarrow \Pi^{-1}(\widehat{x}_k)\widehat{A}^T \left(\beta(\widehat{x}_k) I + \widehat{A}\Pi^{-1}(\widehat{x}_k)\widehat{A}^T\right)^{-1} y_k , \qquad (18)$$

$k = 1, \cdots, N$, with $\Pi(x)$ defined as in equation (24) given below. This is the regularized FOCUSS algorithm described in [18, 2] which has an interpretation as an AST-like concave function minimization algorithm.

The proposed dictionary learning algorithm *alternates* between the interation (18) and the iteration (17) (or the direct batch solution given by (15), if the inversion is tractable). Extensive simulations showing the ability of the AST-based algorithm to completely recover an unknown 20×30 dictionary matrix A can be found in [2].

Unknown, Random Dictionary. We now generalize to the case where the dictionary, A, and the source vector set $X = X^N = (x_1, \cdots, x_N)$ are jointly random. We add the requirement that the dictionary is known to obey the constraint,

$$A \in \mathcal{A} = \text{compact submanifold of } \mathsf{R}^{m\times n}.$$

A compact submanifold of $\mathsf{R}^{m\times n}$ is necessarily closed and bounded. On the constraint submanifold the dictionary A has the prior probability density function $P(A)$, which in the sequel we assume has the simple (uniform on $\mathcal{A}$) form,

$$P(A) = c\,\mathcal{X}(A \in \mathcal{A})\,, \tag{19}$$

where $\mathcal{X}(\cdot)$ is the indicator function and c is a positive constant chosen to ensure that

$$P(\mathcal{A}) = \int_{\mathcal{A}} P(A)\,dA = 1\,.$$

The dictionary A and the elements of the set X are also all assumed to be mutually independent,

$$P(A,X) = P(A)\,P(X) = P(A)P_p(x_1)\cdots P_p(x_N)\,.$$

With the set of iid noise vectors, $(\nu_1, \cdots, \nu_N)$ also taken to be jointly random with, and independent of, A and X, the observation set $Y = Y^N = (y_1, \cdots y_N)$ is assumed to be generated via the model (1). With these assumptions we have

$$\begin{aligned} P(A,X|Y) &= P(Y|A,X)\,P(A,X)/P(Y) \qquad (20)\\ &= c\,\mathcal{X}(A\in\mathcal{A})\,P(Y|A,X)\,P(X)/P(Y)\\ &= \frac{c\,\mathcal{X}(A\in\mathcal{A})}{P(Y)}\prod_{k=1}^{N} P(y_k|A,x_k)\,P_p(x_k)\\ &= \frac{c\,\mathcal{X}(A\in\mathcal{A})}{P(Y)}\prod_{k=1}^{N} P_q(y-Ax_k)\,P_p(x_k)\,, \end{aligned}$$

using the facts that the observations are conditionally independent and $P(y_k|A,X) = P(y_k|A,x_k)$.

The *jointly* Maximum A Posteriori (MAP) estimates

$$(\widehat{A}_{\text{MAP}}, \widehat{X}_{\text{MAP}}) = (\widehat{A}_{\text{MAP}}, \widehat{x}_{1,\text{MAP}}, \cdots, \widehat{x}_{1,\text{MAP}})$$

are found by maximizing *a posteriori* probability density $P(A,X|Y)$ simultaneously with respect to $A \in \mathcal{A}$ and X. This is equivalent to minimizing the negative logarithm of $P(A,X|Y)$, yielding the optimization problem,

$$(\widehat{A}_{\text{MAP}}, \widehat{X}_{\text{MAP}}) = \arg\min_{A\in\mathcal{A},\,X} (d_q(y-Ax) + \lambda\, d_p(x))_N\,. \tag{21}$$

Note that this is a *joint* minimization of the sample average of the functional (5), and as such is a natural generalization of the single (with respect to the set of source vectors) optimization previously encountered in (12). By finding joint MAP estimates of A and X, we obtain a problem that is much more tractable than the one of finding the single MAP estimate of A (which involves maximizing the marginal posterior density $P(A|Y)$).

The requirement that $A \in \mathcal{A}$, where $\mathcal{A}$ is a compact and hence *bounded* subset of $\mathsf{R}^{m\times n}$, is sufficient for the optimization problem (21) to avoid the trivial solution,

$$y_k = Ax_k\,,\ \text{with } \|A\| \to \infty \text{ and } \|x_k\| \to 0\,. \tag{22}$$

This solution is possible for unbounded A because $y = Ax$ is almost always solvable for x since learned overcomplete A's are (generically) onto and for any solution pair (A,x) the pair $(\frac{1}{\alpha}A, \alpha\,x)$ is also a solution. This fact shows that the inverse problem of finding a solution pair (A,x) is generally ill-posed *unless* A is constrained to be bounded (as we've explicitly done here) or the cost functional is chosen to ensure that bounded A's are learned (e.g., by adding a term monotonic in $\|A\|$ to the cost function in (21)).

For the iid $q = 2$ gaussian measurement noise case of (14), algorithms that provably converge (in the low step-size limit) to a local minimum of (21) can be readily developed for the case,

$$\mathcal{A} = \{A \mid \|A\|_F = 1\} \subset \mathsf{R}^{m\times n}\,, \tag{23}$$

where $\|A\|_F$ denotes the Frobenius norm of the matrix A,

$$\|A\|_F^2 = \text{trace}(A^TA)\,,$$

and it is assumed that the prior $P(A)$ is uniformly distributed on $\mathcal{A}$ as per condition (19). Following the procedure described in [5, 17], we factor the gradient of $d(x)$ as

$$\nabla d(x) = \alpha(x)\Pi(x)x\,, \quad \alpha(x) > 0\,, \tag{24}$$

where it is assumed that $\Pi(x)$ is diagonal and positive-definite for all nonzero x. We also define $\beta(x) = \lambda\alpha(x)$. A learning law which provably converges to a minimum of (21) on the manifold (23) is then given by,

$$\begin{aligned} \frac{d}{dt}\widehat{x}_k &= -\Omega_k\left\{\left(\widehat{A}^T\widehat{A} + \beta(\widehat{x}_k)\Pi(\widehat{x}_k)\right)\widehat{x}_k - \widehat{A}^T y_k\right\}\,,\\ \frac{d}{dt}\widehat{A} &= -\alpha\left(\delta\widehat{A} - \text{trace}(\widehat{A}^T\delta\widehat{A})\widehat{A}\right)\,, \end{aligned} \tag{25}$$

for $k = 1, \cdots N$, where $\widehat{A}$ is initialized to $\|\widehat{A}\|_F = 1$, Ω_k are $n \times n$ positive definite matrices, and

$$\delta\widehat{A} = \sum_{k=1}^{N} e(\widehat{x}_k)\widehat{x}_k^T\,, \quad e(\widehat{x}_k) = \widehat{A}\widehat{x}_k - y_k\,. \tag{26}$$

The sign of α must be adaptively chosen as the algorithm progresses, but usually can be taken to be positive with no harm to the stability analysis.

Note that (except for the trace term) the dictionary learning update in (25) is of the same form as for the AML update law given earlier in (17). The key difference is the additional trace term in (25). This difference corresponds to a projection of the update onto the tangent space of the manifold (23), thereby ensuring a unit Frobenius norm (and hence boundedness) of the dictionary estimate at all times and avoiding the ill-posedness problem indicated in (22).

Convergence of the algorithm to a local optimum of (21) is formally proved by interpreting the loss functional as a Lyapunov function whose time derivative along the trajectories of the adapted parameters $(\widehat{A}, \widehat{X})$ is guaranteed to be negative-definite by the choice of parameter time derivatives shown in (25). As a consequence of the La Salle invariance principle, the loss functional will decrease in value and the parameters will converge to the largest invariant set for which the time derivative of the loss functional is identically zero [4].

It is of interest to note that choosing Ω_k to be

$$\Omega_k \propto \left(\widehat{A}^T\widehat{A} + \beta(\widehat{x}_k)\Pi(\widehat{x}_k)\right)^{-1} \tag{27}$$

in (25), followed by some matrix manipulations, yields the alternative algorithm,

$$\frac{d}{dt}\widehat{x}_k = -\eta_k \left\{\widehat{x}_k - \Pi^{-1}(\widehat{x}_k)\widehat{A}^T\left(\beta(\widehat{x}_k)I + \widehat{A}\Pi^{-1}(\widehat{x}_k)\widehat{A}^T\right)^{-1} y_k\right\} \tag{28}$$

with $\eta_k > 0$. In any event (regardless of the specific choice of the postive definite matrices Ω_k) The proposed algorithm outlined here converges to a solution which satisfies the implicit and nonlinear relationships,

$$\begin{aligned} \widehat{x}_k &= \Pi^{-1}(\widehat{x}_k)\widehat{A}^T\left(\beta(\widehat{x}_k)I + \widehat{A}\Pi^{-1}(\widehat{x}_k)\widehat{A}^T\right)^{-1} y_k\,, \\ \widehat{A} &= \Sigma_{y\hat{x}}^T\Sigma_{\hat{x}\hat{x}}^{-1}\,, \end{aligned} \tag{29}$$

for $k = 1, \cdots, N$. When implemented in discrete-time, the Bayesian learning algorithm has the form of a *combined interation* where we loop over the operations,

$$\begin{aligned} \widehat{x}_k &\leftarrow \Pi^{-1}(\widehat{x}_k)\widehat{A}^T\left(\beta(\widehat{x}_k)I + \widehat{A}\Pi^{-1}(\widehat{x}_k)\widehat{A}^T\right)^{-1} y_k\,, \\ k &= 1, \cdots, N \quad \text{and} \\ \widehat{A} &\leftarrow \widehat{A} - \alpha\left(\delta\widehat{A} - \text{trace}(\widehat{A}^T\delta\widehat{A})\widehat{A}\right)\,. \end{aligned} \tag{30}$$

This *merged* procedure should be compared to the *separate* iterations involved in the maximum likelihood approach given in (17)-(18) above. The projection in (30) of the dictionary update onto the tangent plane of $\mathcal{A}$ is critical to ensuring the well-behavedness of the MAP algorithm. Further analysis, discrete-time variants, and simulations of the algorithms proposed in this subsection will be presented in the near future.

References

[1] A. Basilevsky, *Statistical factor analysis and related methods: theory and applications*, Wiley 1994.

[2] K. Engan, B.D. Rao and K. Kreutz-Delgado, "Frame Design Using FOCUSS with Method of Optimal Directions (MOD)", *Proc. NORSIG-99.*

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

[4] H.K. Khalil, *Nonlinear Systems*, 2nd Ed., Prentice Hall, 1996.

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

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

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

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

[9] K. Kreutz-Delgado, B.D. Rao, K. Engan, T.-W. Lee, and T. Sejnowski, "Convex/Schur–Convex (CSC) Log-Priors and Sparse Coding," *Proc. 6th JSNC*, 1999.

[10] K. Kreutz-Delgado, B.D. Rao, K. Engan, T.-W. Lee, and T. Sejnowski, "Learning Overcomplete Dictionaries," *Proc. 6th JSNC*, 1999.

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

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

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

[14] 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).

[15] 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.*

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

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

[18] B.D. Rao, "Signal Processing with the Sparseness Constraint," *Proc. 1998 ICASSP.*