SPARSE BASIS SELECTION, ICA, AND MAJORIZATION: TOWARDS A UNIFIED PERSPECTIVE

K. Kreutz-Delgado and B.D. Rao

Electrical and Computer Engineering
University of California, San Diego
{kkreutzd, brao}@ucsd.edu

Abstract

Sparse solutions to the linear inverse problem $Ax = y$ and the determination of an environmentally adapted overcomplete dictionary (the columns of A) depend upon the choice of a "regularizing function" $d(x)$ in several recently proposed procedures. We discuss the interpretation of $d(x)$ within a Bayesian framework, and the desirable properties that "good" (i.e., sparsity ensuring) regularizing functions, $d(x)$ might have. These properties are: *Schur-concavity* ($d(x)$ is consistent with majorization); *concavity* ($d(x)$ has sparse minima); *parameterizability* ($d(x)$ is drawn from a large, parameterizable class); and *factorizability* of the gradient of $d(x)$ in a certain manner. The last property (which naturally leads one to consider *separable* regularizing functions) allows $d(x)$ to be efficiently minimized subject to $Ax = y$ using an Affine Scaling Transformation (AST)-like algorithm "adapted" to the choice of $d(x)$. A *Bayesian framework* allows the algorithm to be interpreted as an *Independent Component Analysis* (ICA) procedure.

1. INTRODUCTION

Sparsity and Adaptation. For $A = [a_1, \cdots, a_n] \in R^{n \times m} =$ "overcomplete dictionary" ($n > m$, rank$(A) = m$), a *sparse solution*, $\hat{x}$, to the inverse problem $Ax = y$ is a solution having a maximal number of zero elements. Recently, various researchers have noted that sparse solutions can be found found as solutions to a "regularized" inverse problem,

$$\hat{x} = \arg\min_x \left\{ \|Ax - y\|^2 + \gamma\, d(x) \right\}, \tag{1}$$

for an appropriately chosen regularization function $d(x)$ [21, 22, 17, 18, 33, 11]. In the "low noise limit" (see the discussion below), $\gamma \to 0$, this becomes

$$\hat{x} = \min_{Ax=y} d(x). \tag{2}$$

Furthermore, to *learn* a dictionary adapted to the environment, it has been suggested that one choose [33, 11],

$$\hat{A} = \arg \min_{A, x_1, x_2, \cdots} \left\langle \left\{ \|Ax - y\|^2 + \gamma\, d(x) \right\} \right\rangle, \tag{3}$$

where $\langle \cdot \rangle$ indicates an averaging over environmental samples $y \in \{y_1, y_2, \cdots\}$ generated by the "source vectors" $\{x_1, x_2, \cdots\}$. It is evident that choice of $d(x)$ affects the nature of the learned dictionary A and, given A, any particular sparse solution $\hat{x}$.

Bayesian Interpretation. The function $d(x)$ has a statistical interpretation as a (negative logarithm of) a Bayesian prior [17, 18, 21, 22] or (deterministically) as a penalty function enforcing sparsity where $d(x)$ should serve as a "relaxed counting function" on the nonzero elements of x [4, 13, 7]. Our approach emphasizes the fact that $d(x)$ can serve *both* of these roles simultaneously. It may also be useful to have a large parameterized class of functions, $d(x) = d_\lambda(x)$, to allow for the possibility of selecting a "best" $d(x)$ in specific applications. A Bayesian interpretation is obtained from the generative signal model,

$$y = Ax + \nu, \tag{4}$$

where x has the parameterized pdf,

$$p_\lambda(x) = Z_\lambda^{-1} e^{-d_\lambda(x)}, \quad Z_\lambda = \int e^{-d_\lambda(x)} dx,$$

and ν is assumed to be normally distributed, $p_\nu \sim \mathrm{N}(0, \frac{\gamma}{2} \cdot I)$. Treating A deterministically (this can be relaxed), but perhaps unknown, and assuming that x and ν are independent (and that the parameters λ and A have no functional dependencies), Bayes' rule yields,

$$p(s|y; \lambda, A) = \frac{1}{\beta} p(y|s; \lambda, A) \cdot p(s; \lambda, A) = \frac{1}{\beta} p_\nu(y - As) \cdot p_\lambda(s),$$

where

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

When the "prior" $p_\lambda(s)$ and the dictionary A are both *pre-specified*, maximizing this expression (equivalently, minimizing its negative logarithm) leads to the optimization problem (1) discussed above. This results in a *maximum a posteriori* estimate, $\hat{x}$ of the generating signal vector x given the observed signal y. We have studied this case in the low noise limit [23, 8], and more recently we have begun to consider algorithms appropriate when noise is non-negligible [26].

For *unknown* λ and/or A, it is required to learn values of λ and 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 (4) that best explains the observations. Note that the source vector x is unknown ("blind source problem") which makes the task of estimating λ 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 we can obtain *maximum likelihood* estimates of λ, A, and X^N given Y^n by solving the problem

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

This can be written in terms of the sample average as

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

which with the generative model (4) and fixed λ gives the optimization problem (3) mentioned above. This approach is analyzed for the case of the ℓ_1 norm prior in [17, 18, 11, 16].

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 [15, 2, 19, 20, 27]. Ideally, optimizing over the parameterization λ will provide a good parametric fit of $p_\lambda(x)$ to the true underlying environmental prior probability density function describing the source vectors. Thus we see the potential desirability of having a relatively large family of distributions $p_\lambda(x)$. The problem of selecting 'an 'optimal" choice of λ is known as the problem of hyperparameter selection in the Bayesian estimation literature [32, 33].

As previously mentioned, the functions $d_\lambda(x)$ should also have analytical properties consistent with the goal of enforcing sparse solutions. In our recent work we have focused on the second desirable aspect of $d(x)$, i.e. on the requirement that $d(x)$ be sparsity-enforcing, and on the development of algorithms to solve the low-noise problem (2) given a specified overcomplete dictionary A [23, 8]. We discuss this aspect in more detail next.

2. MAJORIZATION AND SCHUR-CONCAVITY

Schur-concave functions. A measure of the sparsity of the elements of a solution vector x, or the lack thereof (which we refer to as the *diversity* of x) is given by a partial ordering on vectors known as the *Lorentz order*. For any vector in the positive orthant, $x \in R_+^n$, define the *decreasing rearrangement*

$$x \doteq (x_{\lfloor 1\rfloor}, \cdots, x_{\lfloor n\rfloor})\,, \quad x_{\lfloor 1\rfloor} \geq \cdots \geq x_{\lfloor n\rfloor} \geq 0$$

and the *partial sums* [28,16],

$$S_x[k] = \sum_{i=1}^{k} x_{\lfloor i\rfloor}\,, \quad k = 1, \cdots, n\,.$$

We say that y *majorizes* x, $y \succ x$, iff for $k = 1, \cdots, n$,

$$S_y[k] \geq S_x[k]\,; \quad S_y[n] = S_x[n]\,.$$

The vector y is more concentrated, or less *diverse*, than x. This partial order defined by majorization defines the Lorentz order.

We are interested in scalar-valued functions of x which are consistent with majorization. These are known as *Schur-Concave* functions, $d(\cdot) : R_+^n \to R$. They are defined to be precisely the class of functions which are *consistent with the Lorentz order*,

$$y \succ x \quad \Rightarrow \quad d(y) < d(x)\,.$$

In words, if y is *less diverse than* x (according to the Lorentz order) then $d(y)$ is *less than* $d(x)$ for $d(\cdot)$ Schur-concave. Henceforth we take Schur-Concavity to be a *necessary condition* for $d(\cdot)$ to be a good *measure of diversity* (*anti-sparsity*).

Concavity yields sparse solutions. Recall that a function $d(\cdot)$ is *concave* on the positive orthant R_+^n iff [28],

$$d\left((1-\gamma)x + \gamma y\right) \geq (1-\gamma)d(x) + \gamma d(y)\,,$$

$\forall x, y \in R_+^n, \forall \gamma, 0 \leq \gamma \leq 1$. A scalar function is said to be *permutation invariant* or *symmetric* if its value is independent of rearrangements of its components. An important fact is that for permutation invariant functions *concavity is a sufficient condition for Schur-Concavity* [16]:

$$\text{Concavity + Symmetry } \Rightarrow \text{ Schur-Concavity.}$$

Now recall the low-noise sparse inverse problem (2). It is well known that subject to linear constraints, a concave function on R_+^n takes its minima on the *boundary* of R_+^n [28], and as a consequence these minima are therefore *sparse*. We take concavity to be a *sufficient condition* for $d(\cdot)$ to be a measure of diversity and we obtain sparsity as constrained minima of $d(\cdot)$.

More generally, a diversity measure should be somewhere between Schur-concave and concave. In [8] are defined *almost concave* functions, which are Schur-concave and (locally) concave in all n directions but one, which also are good measures of diversity.

Separability, Schur-Concavity, and ICA. The simplest way to ensure that $d(x)$ be permutation invariant (a necessary condition for Schur-concavity) is to use functions that are *separable*. Separable functions 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 $p_\lambda(x)$,

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

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) [15, 2, 19, 20, 27].

It is now evident that relaxing the restriction of separability generalizes the generative model to the case were 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). Unfortunately, this relaxation appears to significantly complicate the analysis and development of optimization algorithms.

3. ADMISSIBLE DIVERSITY MEASURES

Separable Measures. The diversities measures considered in [23, 8] are separable, with only a brief mention of the extension to non-separable measures given in [8]. In addition to separable measures based on the Shannon entropy function, we have considered in some detail the following functions.

P-Class. We define $d_p(x) = \mathrm{sgn}(p)\sum_{i=1}^{n}|x[i]|^p\,, \quad p \le 1$. The separable p-class generalizes the ℓ_1-norm measures to $p \le 1$, including p negative. Every p-class function (excluding $p \equiv 0$) is concave and permutation invariant, and hence Schur-concave.The ℓ_1-norm case, $p = 1$, corresponds to the choice of a exponential density function as a Bayesian prior; a simple choice that is often used in the ICA literature [17, 18]. Note that $p = 2$ is *not* concave and is not a p-class function. The case $p = 2$ corresponds to assuming a gaussian (maximum entropy) prior on solutions x and results in *nonsparse* solutions to (2).

$\mathcal{S}$-functions. This large parameterized class of separable permutation invariant diversity measures is a superset of the p-class discussed earlier and is defined by [8],

$$\begin{aligned} d_{\mathcal{S}}(x) &= \sum_{i=1}^{n}\mathsf{S}(|x[i]|) = \sum_{j=1}^{q}\omega_j\, d_{p_j}(x)\,, \\ d_{p_j}(x) &= \mathrm{sgn}(p_j)\sum_{i=1}^{n}|x[i]|^{p_j}\,, \quad p_j \le 1\,, \qquad (5) \\ \mathsf{S}(s) &= \mathrm{sgn}(p_1)\,\omega_1\, s^{p_1} + \cdots + \mathrm{sgn}(p_q)\,\omega_q\, s^{p_q}, \\ s > 0\,,\ p_j &< 1\,, \quad p_j \neq 0\,, \quad \text{and} \quad \omega_j \ge 0\,, \\ \text{or} \quad p_j &= 0, 1\,, \quad \text{and} \quad \omega_j \in \mathsf{R}\,. \end{aligned}$$

Note that the $\mathcal{S}$-functions have fractional and possibly negative powers, $p_j \le 1$ and are strictly concave. The $\mathcal{S}$-functions *provide a rich class of regularizing functions* for the functional (1) which can be used to affect the nature of the basis vectors of a dictionary constructed from optimizing (2) with respect to A over an ensemble of environmental samples. A natural question to ask is if this parameterized class can prove useful for obtaining good (factorizable) probability density function estimates for the prior density of the source vector x.

Nonseparable $\mathcal{S}$-functions. As proved in [14], every permutation invariant concave function is Schur-concave. To generalize the separable $\mathcal{S}$-functions defined above to include nonseparable functions requires that we include symmetric "cross terms" of sums of products of powers of the components of x and give conditions to ensure concavity of the resulting permutation symmetric functions. Alternatively, we can define a non-symmetric concave function of the components of x and proceed to symmetrize it using the methodologies discussed in [14]. In any event, we find that the problem of producing symmetric, concave diversity measures having a simple structure for the gradient factorization discussed below is significantly more complicated.

4. AFFINE SCALING AND GRADIENT FACTORIZATION

An AST-Like Algorithm. In [23, 8], we have shown that diversity measures $d(x)$ can be efficiently minimized subject to $Ax = y$ using an Affine Scaling Transformation (AST)-like algorithm which is "adapted" to the choice of $d(x)$. Towards this end, the *gradient* of an admissible diversity measure $d(x)$ is *factored* as

$$\nabla d(x) = \alpha(x)\Pi(x)x\,, \qquad (6)$$

where $\alpha(x)$ is a positive scalar function, and $\Pi(x)$ is the *scaling matrix*. The quantities $\alpha(x)$ and $\Pi(x)$ are invariant with respect to permutations of the elements of x. The scaling matrix $\Pi(x)$ and its properties turns out to be key in constructing a convergent, recursive algorithm that will provably converge to a local minimum (and therefore sparse solution) of the problem (2). A *positive definite* scaling matrix $\Pi(x)$ defines a *natural* ($d(x)$-dependent) Affine Scaling Transformation (AST) matrix $W(x)$ by

$$W(x) \triangleq \Pi^{-\frac{1}{2}}(x). \qquad (7)$$

The relevant result concerning the behavior of the algorithm is given by the following theorem which can be proved using the general convergence theorem of Zangwill [31].

Theorem 1 ([8]) *Let $d(x)$ be a sign and permutation invariant function that is strictly concave on the positive orthant and for which $\Pi(x) > 0$ for all $x \in \mathsf{R}^n$. Assume that the set $\{x|\, d(x) \le d(x_0)\}$ is compact for all x_0. Let x_k be generated by the algorithm*

$$\begin{aligned} W_{k+1} &= W(x_k)\,, \quad A_{k+1} = AW_{k+1}\,, \qquad (8) \\ q_{k+1} &= A_{k+1}^{+}y\,, \quad x_{k+1} = W_{k+1}q_{k+1}\,, \qquad (9) \end{aligned}$$

with A_{k+1}^{+} the Moore-Penrose pseudoinverse of A, starting with x_0 feasible, $Ax_0 = y$. Then for all $|x_{k+1}| \neq |x_k|$ (the function $|\cdot|$ is defined component-wise), we have x_k is feasible, $d(x_{k+1}) < d(x_k)$, and the algorithm converges to a local minimum $d(x^*)$, $x_k \to x^*$, where x^* is a boundary point of some orthant and $Ax^* = y$.

As discussed in [23, 8], and mentioned above, convergence to a boundary point ensures that a sparse solution to the inverse problem $Ax = y$, although in general the solution will be a local, but not a global, solution to the problem (2).

Separable Diversity Measures. For the *separable* $\mathcal{S}$-function diversity measures the scaling matrix Π has a simple *diagonal* form (and therefore easily invertible) and is positive definite. For $d_{\mathcal{S}}(x)$ an $\mathcal{S}$-function we have

$$\Pi(x) = \sum_j |p_j|\,\omega_j\,\Pi_{p_j}(x)\,; \quad \Pi_{p_j}(x) = \mathrm{diag}\left(\frac{1}{|x[i]|^{2-p_j}}\right).$$

This yields a rich class of *separable* diversity measures which satisfy the conditions of Theorem 1 and which can be used to solve the low-noise problem (2). Utilization of the algorithm for functions drawn from the $\mathcal{S}$-class (which contains the p-class) will result in sparse solutions to $Ax = y$ for specified $d_{\mathcal{S}}(x)$ and dictionary A. Thus we have an algorithm that can provide sparse solutions to the ICA problem for factorizable priors drawn from the $\mathcal{S}$-class of diversity measures (and given a dictionary A). Convergence of the algorithm for other separable diversity measures, such as the Shannon and Gaussian measures, is discussed in [23, 8].

5. DISCUSSION AND CONCLUSIONS

The algorithm of Theorem 1 was originally derived for constrained minimization of the (separable) Gaussian entropy in [5, 6]. Properly interpreted within the majorization/concavity framework, reference [6] provides a rigorous justification for the use of the p-class of diversity measures, d_p for $p \ge 0$. In [23] it was shown that the algorithm derived for the Gaussian measure $d_G(x)$ and that for the limiting p-class measure $\lim_{p\to 0} d_p(x)$ are identical and the relationship between these two separable measures was examined. Convergence was also shown for the Shannon entropy measures. A general analysis of the case of positive definite scaling matrix

$\Pi(x)$ (which includes the separable general $\mathcal{S}$-class of measures) and other cases (including Renyi entropy-based measures) can be found in [8]. An interesting application to MEG signal processing is given in [5, 6]. Other references are [1, 24, 25, 26, 9, 10].

Obviously, much work remains to be done. We are currently investigating the use of multiple measurements for enhanced noise robustness, extending the algorithm to the "noisy" case (1), and considering extensions of the algorithm to efficiently solve the environmentally adapted basis learning problem (3). We are also continuing to examine the problem of adapting the parameters (known as hyperparameters in the Bayesian estimation and learning literature [32]) of the $\mathcal{S}$-class of diversity functions to provide a good density estimate, $p_\lambda(x)$ of the unknown density of x. In addition we are looking into rate-of-convergence and scaling issues associated with our algorithm.

The majorization framework, focusing on sparse solution requirements, complements the Bayesian/ICA framework, which is concerned with statistically sound solutions to the signal representation problem. We have seen that the separable, Schur-concave/concave diversity measures are particularly interesting in that they lead to straightforward algorithm development (at least in the low-noise case) and correspond to the use of a factorizable prior appropriate for obtaining ICA solutions [15, 27]. In particular, the $\mathcal{S}$-class of diversity measures provides a large, parameterized class of separable diversity measures. In principle, a measure can be drawn from this class in an "optimal" manner to better model the *a priori* statistical properties of the environment, allowing for a parametric density estimation of the environmental Bayes prior subject to the constraint that the negative logarithm of the prior is concave.

6. REFERENCES

[1] J.M. Adler and B.D. Rao, "Comparison of Basis Selection Methods", *Proc. 30th Asilomar Conference*, November 1996.

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

[3] R.R. Coifman and M.V. Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection", *IEEE Transactions on Information Theory*, IT-38(2):713–18, March 1992.

[4] D. Donoho, "On Minimum Entropy Segmentation", C.K. Chui, L. Montefusco, and L. Puccio, editors, *Wavelets: Theory, Algorithms, and Applications*, p. 233-69. AP, 1994.

[5] I.F. Gorodnitsky, J.S. George, and B.D. Rao, "Neuromagnetic Source Imaging with FOCUSS: a Recursive Weighted Minimum Norm Algorithm", *Journal of Elect. Clin. Neurophysiology*, 95(4):231-51, 1995.

[6] I.F. Gorodnitsky and B.D. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm', *IEEE Trans. Signal Processing*, 45(3):600-16, 1997.

[7] G. Harikumar and Y. Bresler,"A New Algorithm for Computing Sparse Solutions to Linear Inverse Problems", *Proc. ICASSP 96*, volume III, p. 1331-4, 1996.

[8] K. Kreutz-Delgado and B.D. Rao, *A General Approach to Sparse Basis Selection: Majorization, Concavity, and Affine Scaling*, Report No. UCSD-CIE-97-7-1, ECE Department, UCSD 1997.

[9] K. Kreutz-Delgado and B.D. Rao, "Gradient Factorization-Based Algorithm for Best-Basis Selection", *Proc. 8th IEEE Digital Signal Processing Workshop*, 1998.

[10] K. Kreutz-Delgado and B.D. Rao, "Measures and Algorithms for Best Basis Selection", *Proc. 1998 ICASSP*, New York, 1998.

[11] M. Lewicki and T.J Sejnowski, "Learning Overcomplete Representations", February 1998. Submitted to *Neural Computation.*

[12] M. Malfait and D. Roose, "Wavelet-Based Image Denoising Using a Markov Random Field a Priori Model," *IEEE Trans. Image Proc.*, 6(4):549-65, 1997.

[13] O.L. Mangasarian, *Machine Learning via Polyhedral Concave Minimization*, November 1995, Mathematical Programming Technical Report 95-20, Computer Sciences Department, University of Wisconsin.

[14] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. Academic Press, 1979.

[15] 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.

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

[17] 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.

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

[19] 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).

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

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

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

[23] B.D. Rao and K. Kreutz-Delgado, *An Affine Scaling Methodology for Best Basis Selection*, 1997. To appear in *IEEE Trans. on Signal Processing.*

[24] B.D. Rao and K. Kreutz-Delgado, "Deriving Algorithms for Computing Sparse Solutions to Linear Inverse Problems", In J. Neyman, editor, *Proc. 1997 Asilomar Conf. Circuits, Sys. Computers*, 1997.

[25] B.D. Rao and K. Kreutz-Delgado,"Sparse Solutions to Linear Inverse Problems with Multiple Measurement Vectors", *Proc. 8th IEEE Digital Signal Processing Workshop*, 1998.

[26] B.D. Rao, K. Kreutz-Delgado, and S. Dharanipragada, "Improving Spectral Resolution Using Basis Selection Methods", *Proc. 9th IEEE Signal Pro. Workshop on Statist. Signal and Array Proc.*, 1998.

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

[28] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

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

[30] M.V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A.K. Peters, 1994.

[31] W.I. Zangwill, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, 1969.

[32] 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.

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