MEASURES AND ALGORITHMS FOR BEST BASIS SELECTION

K. Kreutz-Delgado and B.D. Rao

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

ABSTRACT

A general framework based on majorization, Schur-concavity, and concavity is given that facilitates the analysis of algorithm performance and clarifies the relationships between existing proposed diversity measures useful for best basis selection. Admissible sparsity measures are given by the Schur-concave functions, which are the class of functions consistent with the partial ordering on vectors known as majorization. Concave functions form an important subclass of the Schur-concave functions which attain their minima at sparse solutions to the basis selection problem. Based on a particular functional factorization of the gradient, we give a general affine scaling optimization algorithm that converges to a sparse solution for measures chosen from within this subclass.

1. INTRODUCTION

There has been considerable recent interest in the issue of best basis selection for sparse signal representation, including approaches that select basis vectors by minimizing diversity measures subject to the constraint

$$Ax = b, \tag{1}$$

where A is an $m \times n$ matrix formed using the vectors from an overdetermined dictionary of basis vectors, $m < n$, and it is assumed that $\text{rank}(A) = m$ [3, 13, 1, 10].

The system of equations (1) has infinitely many solutions, and the solution set is a linear variety denoted by $LV(A,b) = x_p + \mathcal{N}(A)$, where x_p is any particular solution to (1) and $\mathcal{N}(A) =$ Nullspace of A. Constrained minimization of diversity measures results in sparse solutions consistent with membership in $LV(A,b)$. Sparse solutions refer to *basic solutions*, solutions with m nonzero entries, and *degenerate* basic solutions, solutions with less than m nonzero entries [5].

The degenerate basic solutions, if they exist, are desirable from a sparsity objective. The nonzero entries of a sparse solution indicate the basis vectors (columns of A) selected. Popular diversity measures used in this context are the Shannon Entropy, the Gaussian Entropy, and the $\ell_{(p\leq 1)}$ (p-norm-like) diversity measures, $p \leq 1$ [3, 13, 4, 9]. In [13], the Shannon entropy and the $\ell_{(p\leq 1)}$, $0 < p \leq 1$, measures, both evaluated on the "probability" $\tilde{x} = |x|^2/\|x\|_2^2 \in \mathbb{R}^n$, are analyzed at length.[1] It is shown that these functions are consistent with diversity as measured by the partial sums of the decreasing rearrangement of the elements of $\tilde{x}$. Ordering of vectors according to their partial sums is known as *majorization* and many results relating majorization to functional inequalities exist that can be exploited to more fully understand the relationship between majorization and measures of diversity [2, 7].

Inspired by the insightful discussion given in Chapter 8 of [13], we have been motivated to analyze and develop diversity measures from the perspective of majorization theory and to consider measures drawn from the general class of Schur-concave functions, which are precisely the functions consistent with the partial order induced by majorization. In this paper, we argue that diversity measures should be drawn from the class of Schur-concave functions [7] and, in particular, that good diversity measures are a subclass of concave functions. Proofs of the theorems can be found in [12, 7, 9, 6].

[1] We use the notation where $|x|$, x^2, $x^{\frac{1}{2}}$, $x \geq 0$, etc., are defined component-wise for $x \in \mathbb{R}^n$.

2. THE MEASUREMENT OF DIVERSITY

2.1. Majorization and Schur-Concavity

To simplify the discussion, in this section we restrict our discussion to the positive orthant $\mathcal{Q}_1 \subset \mathbb{R}^n$. Let $x_{\lfloor 1 \rfloor} \geq \cdots \geq x_{\lfloor n \rfloor}$ denote the *decreasing rearrangement* of the elements of x and define the sequence of partial sums [13],

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

Definition 1 (Majorization of x by y)

$$x \prec y \quad \text{iff} \quad S_x[k] \leq S_y[k]\,, \quad S_x[n] = S_y[n]\,. \tag{2}$$

When $x \prec y$, then y *majorizes* x (equivalently, x is majorized by y). Often $S_x[k]$ is normalized to one, $S_x[n] = 1$.

A plot of $S_x[k]$ versus k is known as a *Lorentz curve* [7], $\mathcal{L}_x$, and $x \prec y$ iff $\mathcal{L}_y$ is everywhere above the curve $\mathcal{L}_x$. When $x \prec y$, the curve $\mathcal{L}_x$ graphically shows greater equality, or diversity, for the values of the elements of $x \in \mathcal{Q}_1$ than is the case for $\mathcal{L}_y$. The elements of y are more concentrated in value, or less diverse, than the elements of x. This graphical representation explains why majorization is also known as the Lorentz order. Lorentz curves that intersect other than at an end-point correspond to vectors that cannot be ordered by majorization.

When $x \prec y$, we say that x is less concentrated (more diverse) than y or, equivalently, that y is more concentrated than x. It is natural to ask what functions $\phi(\cdot)$ are consistent with the diversity ordering provided by majorization.

0-7803-4428-6/98 $10.00 © 1998 IEEE

Definition 2 *A function ϕ is called* permutation invariant *iff it is invariant with respect to all permutations of its argument x, i.e. $\phi(x) = \phi(Px)$ for any permutation matrix P.*

Definition 3 *A function $\phi : \mathrm{R}^n \to \mathrm{R}$ is said to be* Schur-concave *if $\phi(x) \geq \phi(y)$ whenever $x \prec y$, and* strictly *Schur-concave if in addition $\phi(x) > \phi(y)$ when x is also not a permutation of y.*

A Schur-concave function is necessarily invariant with respect to permutations of the elements of x. For $\phi(\cdot)$ Schur-concave, it is natural to consider x to be more diverse than y, if $\phi(x) \geq \phi(y)$ [7, 8]. An approach to sparse basis selection can then be based on minimizing diversity, as measured by a Schur-concave function $\phi(\cdot)$, subject to the constraint (1).

Theorem 1 *A function ϕ is Schur-concave on $\mathcal{Q}_1$ iff it is permutation invariant and satisfies the* Schur condition,

$$(x[i] - x[j]) \left(\frac{\partial \phi(x)}{\partial x[i]} - \frac{\partial \phi(x)}{\partial x[j]} \right) \leq 0\,, \quad \forall x \in \mathcal{Q}_1\,. \quad (3)$$

Furthermore, because of the assumed permutation invariance of $\phi(x)$, one only need verify (3) for a single set of specific values for the pair (i, j).

Theorem 2 *If $\phi(\cdot)$ is Schur-concave on the interior of $\mathcal{Q}_1$, then the scale invariant function ψ defined by $\psi(x) = \phi(x/\|x\|_1)$ is also Schur-concave on the interior of $\mathcal{Q}_1$.*

2.2. Concave Functions as Measures of Diversity

Theorem 3 *Let x, y belong to a permutation symmetric, convex set $\mathcal{C} \subset \mathrm{R}^n$. Then $x \prec y$ iff $\phi(x) \geq \phi(y)$ for all permutation invariant and concave functions $\phi : \mathcal{C} \to \mathrm{R}$.*

A particularly useful and tractable set of diversity measures is provided by the subclass of separable concave functions.

Definition 4 *A function $\phi : \mathrm{R}^n \to \mathrm{R}$ is* separable *if there exists $g : \mathrm{R} \to \mathrm{R}$ such that $\phi(x) = \sum_{i=1}^n g(x[i])$.*

Theorem 4 *Let x, y belong to a permutation symmetric, convex set $\mathcal{C} \subset \mathrm{R}^n$. Then $x \prec y$ iff $\sum_{i=1}^n g(x[i]) \geq \sum_{i=1}^n g(y[i])$ for every concave function $g : \mathcal{C} \to \mathrm{R}$.*

Theorem 5 *Let $\phi : \mathcal{C} \to \mathrm{R}$ be strictly concave and bounded from below on a closed convex set $\mathcal{C} \subset \mathrm{R}^n$. Then ϕ attains its local minima (and hence its global minima) at boundary points of $\mathcal{C}$.*

Theorem 6 *Let $\phi : \mathcal{C} \to \mathrm{R}$ be concave on a closed convex set $\mathcal{C} \subset \mathrm{R}^n$ which contains no lines. If ϕ attains a global minimum somewhere on $\mathcal{C}$, it is also attained at an extreme point of $\mathcal{C}$.*

Definition 5 *A function ϕ is said to be* sign invariant *if $\phi(x) = \phi(\bar{x}), \forall x \in \mathrm{R}^n$, where $\bar{x} = |x| \in \mathcal{Q}_1$.*

Theorem 7 *Let $\phi : \mathrm{R}^n \to \mathrm{R}$ be permutation invariant, sign invariant, and concave on the positive orthant $\mathcal{Q}_1$. Then the global minimum of $\phi(x)$ subject to the linear constraints of (1) is attained at a basic solution.*

Theorems 3–7 show that the permutation and sign invariant concave functions are particularly good measures of diversity, if our intent is to obtain sparse solutions to (1) by minimizing diversity measures. The following two theorems can be used to identify concave diversity measures.

Theorem 8 *Let $\mathcal{C} \subset \mathrm{R}^n$ be an open convex set and let $\phi : \mathcal{C} \to \mathrm{R}$ be differentiable on $\mathcal{C}$. Then ϕ is concave on $\mathcal{C}$ iff for any $x \in \mathcal{C}$ we have*

$$\nabla \phi(x)^T (y - x) \geq \phi(x) - \phi(y)\,, \quad \forall y \in \mathcal{C}\,. \quad (4)$$

Furthermore ϕ is strictly concave iff the inequality is strict for every $y \neq x$.

Theorem 9 *Let $\mathcal{C} \subset \mathrm{R}^n$ be an open convex set and let $\phi : \mathcal{C} \to \mathrm{R}$ be twice differentiable on $\mathcal{C}$. Let $H(x)$ denote the Hessian matrix of second partial derivatives of ϕ evaluated at the point $x \in \mathcal{C}$. The function ϕ is concave on $\mathcal{C}$ iff for any $x \in \mathcal{C}$ $H(x)$ is negative semidefinite. Furthermore ϕ is strictly concave on $\mathcal{C}$ if $H(x)$ is negative definite for all $x \in \mathcal{C}$.*

3. SCALAR MEASURES OF DIVERSITY

A general diversity measure is henceforth denoted by $d(\cdot) : \mathrm{R}^n \to \mathrm{R}$, and is assumed to be both permutation and sign invariant. Because of the assumed sign invariance, $d(x) = d(|x|)$, Schur-concavity (or concavity) over $\mathcal{Q}_1$ corresponds to Schur-concavity (or, respectively, concavity) over any other orthant $\mathcal{Q}_l$. However, that this does not guarantee Schur-concavity or concavity *across* orthants, and in general this property will not be true.

3.1. Signomial Measures

$\mathcal{S}$-functions. Here, we present a general class of separable concave (and hence Schur-concave) functions that include as a special case the class of $\ell_{(p\leq 1)}$ diversity measures defined by [13, 4, 9],

$$d_p(x) = \mathrm{sgn}(p) \sum_{i=1}^n |x[i]|^p\,, \quad p \leq 1\,. \quad (5)$$

The generalization we are interested in is the subclass of *signomials* given by the separable function,

$$\begin{aligned} d_{\mathrm{sig}}(x) &= \sum_{i=1}^n \mathsf{S}(|x[i]|) = \sum_{j=1}^q \omega_j\, d_{p_j}(x)\,, \quad (6) \\ \mathsf{S}(s) &= \mathrm{sgn}(p_1)\, \omega_1\, s^{p_1} + \cdots + \mathrm{sgn}(p_q)\, \omega_q\, s^{p_q}\,, \\ \text{where } p_j &< 1\,, \quad p_j \neq 0\,, \quad \text{and} \quad \omega_j \geq 0\,, \\ \text{or } p_j &= 0, 1\,, \quad \text{and} \quad \omega_j \in \mathrm{R}\,. \end{aligned}$$

Unlike a regular polynomial, $\mathsf{S}(s)$ has fractional and possibly negative powers, $p_j \leq 1$. With no loss of generality, in (6) we can take $\sum_j \omega_j = 1$.

We will refer to functions of the form (6) as *$\mathcal{S}$-functions.* It is readily shown that $d_{\mathrm{sig}}(x)$ has a diagonal, negative semidefinite Hessian for $x \in \mathcal{Q}_1$. Therefore, from Theorem 9 we know that $d_{\mathrm{sig}}(x)$ is concave on the interior of the positive orthant $\mathcal{Q}_1 \subset \mathrm{R}^n$. Furthermore, if there exists j such that $p_j < 1$, $p_j \neq 0$, then the Hessian is negative definite and $d_{\mathrm{sig}}(x)$ is strictly concave on the interior of the positive orthant $\mathcal{Q}_1$. By construction, $d_{\mathrm{sig}}(x)$ is separable and both sign and permutation invariant (and thus Schur-concave). Furthermore, $d_{\text{sig}}(x)$ can be designed to be strictly concave, insuring that a sparse solution can be obtained to the basis selection problem by searching for a minimum of $d_{\text{sig}}(x)$. Summarizing our results, we have the following theorems.

Theorem 10 *Let x, y belong to a symmetric, convex set $\mathcal{C} \subset \mathcal{Q}_1$. Then $x \prec y$ only if $d_{\text{sig}}(x) \geq d_{\text{sig}}(y)$ for every $\mathcal{S}$-function $d_{\text{sig}} : \mathcal{C} \to \mathrm{R}$.*

Theorem 11 *Every $\mathcal{S}$-function d_{sig} is concave on the interior of $\mathcal{Q}_1$. Furthermore, any $\mathcal{S}$-function such that there exists j for which $p_j < 1$, $p_j \neq 0$, is strictly concave on the interior of $\mathcal{Q}_1$.*

For $p > 1$, $d_p(x)$ of (5) is *not* Schur-concave and hence not concave. Indeed, it is well known (and readily demonstrated) that $d_p(x)$ is *convex* over $\mathcal{Q}_1$ for $p > 1$.

Normalized $\mathcal{S}$-functions. From the class of $\mathcal{S}$-functions, one can define the *1- and 2-normalized $\mathcal{S}$-functions* by taking

$$d^{(1)}_{\text{sig}}(x) = d_{\text{sig}}(\tilde{x}), \quad \tilde{x} = |x|/\|x\|_1, \tag{7}$$

$$d^{(2)}_{\text{sig}}(x) = d_{\text{sig}}(\tilde{x}), \quad \tilde{x} = |x|^2/\|x\|_2^2. \tag{8}$$

In [6] it is shown that with appropriate restrictions on the values of the powers p_j these measures are Schur-concave, but not (quite) concave, functions of x. A slightly weaker property than concavity, *almost* concavity, is defined in [6] and conditions are given that ensure that the normalized $\mathcal{S}$-functions are almost concave.

3.2. Entropy Measures

Gaussian Entropy. Reference [13] proposes the use of the "logarithm of energy" function,

$$H_G(x) = \sum_{i=1}^{n} \log |x[i]|^2, \tag{9}$$

as a measure of diversity and points out that this can be interpreted as the entropy of a Gauss-Markov process; for this reason we refer to (9) as the *Gaussian entropy* measure of diversity. It is straightforward to demonstrate that the Hessian of H_G is everywhere positive definite on the positive orthant $\mathcal{Q}_1$, showing that H_G is strictly concave on the interior of $\mathcal{Q}_1$ and hence Schur-concave. The Gaussian entropy is therefore a good measure of diversity and we expect that minimizing H_G will result in sparse solutions to the best basis selection problem.

In [9], an algorithm is presented to minimize H_G that indeed shows very good performance in obtaining sparse solutions. It is also shown that the algorithm to minimize H_G is the same as the algorithm that minimizes (5) for $p = 0$ and can therefore be given the interpretation of optimizing the numerosity ($p = 0$) measure described by [4]. The interpretation of H_G as a $p = 0$ measure follows naturally from the literature on inequalities where $\mathrm{Exp}(H_G) = (\prod_i |x[i]|)^2$ is shown to be intimately related to the $p = 0$ norm [2, 9].

Shannon Entropy. References [3, 13, 4] have proposed the use of the Shannon entropy function as a measure of diversity appropriate for sparse basis selection. Given a probability distribution, the Shannon entropy is well defined. However starting from x, there is some freedom in how one precisely defines this measure. Defining the Shannon entropy function $H_S(x)$ by

$$H_S(x) = -\sum_{i=1}^{n} \tilde{x}[i] \log \tilde{x}[i], \quad \tilde{x} = \tilde{x}(x) \geq 0, \tag{10}$$

the differences arise in how one defines $\tilde{x}$ as a function of x. These differences affect the properties of H_S as a function of x. It is well known that $H_S(\cdot)$ defined as a function of $\tilde{x}$ by (10) is Schur-concave [7, 13]. However it is generally *not* the case that $H_S(x)$ is Schur-concave with respect to x [7]. In [6] the possible choices $\tilde{x} = |x|$, $\tilde{x} = |x|/\|x\|_1$, and $\tilde{x} = x^2/\|x\|_2^2$ are considered. Whereas the first choice can be shown to result in strict concavity on the interior of $\mathcal{Q}_1$, the second choice results in an almost concave function (in the sense defined in [6]) while the last choice of $\tilde{x}[i] = x^2/\|x\|_2^2$ is not even Schur-concave in x over $\mathcal{Q}_1$.

Renyi Entropy. A family of entropies, parameterized by p, is described in [11]. These *Renyi entropies* include, as a special case, the Shannon entropy. Given a "probability" $\tilde{x}(x)$, $\tilde{x}[i] \geq 0$, $\sum_i \tilde{x}[i] = 1$, the Renyi entropy is defined for $0 \leq p$ by

$$H_p(x) = \frac{1}{1-p} \log \sum_{i=1}^{n} \tilde{x}[i]^p = \frac{1}{1-p} \log d_p(\tilde{x}), \tag{11}$$

where

$$H_1(x) = \lim_{p \to 1} H_p(\tilde{x}) = -\sum_{i=1}^{n} \tilde{x} \log \tilde{x} = H_S(\tilde{x}),$$

is the Shannon entropy of $\tilde{x}$. Because $\log(\cdot)$ is monotonic, we see that for purposes of optimization $H_p(\tilde{x})$ is equivalent to $d_p(\tilde{x})$, and hence is related to the normalized $\mathcal{S}$-functions mentioned above. Thus, consistent with the discussion given in [4], one can also reasonably refer to the normalized p-norm-like measures $\ell_{(p \leq 1)}$ as entropies.

It can be shown that $H_p(|x|/\|x\|_1)$ for $0 < p < 1$ is almost concave (in the sense of [6]) as a consequence of almost concavity of $d^{(1)}_p(x)$ for $0 < p < 1$ and the fact that $\log(\cdot)$ is an increasing concave function. For $p > 1$, $H_p(|x|/\|x\|_1)$ is not even Schur-concave. Similarly, $H_p(|x|^2/\|x\|_2^2)$ is almost concave for $0 \leq p \leq \frac{1}{2}$ and not Schur-concave for $p > 1/2$ (showing that $H_1(|x|^2/\|x\|_2^2)$ is not Schur-concave).

4. SPARSE BASIS SELECTION

To minimize the general classes of concave diversity measures developed in this paper, we can extend the gradient factorization-based methodology described in [9] and develop iterative algorithms which converge to a basic solution of (1) [6].

4.1. The Scaling Matrix $\Pi(x)$

A particular *factored* functional form for the gradient of the diversity measure $d(x)$ with respect to x is essential for the development of the algorithms,

$$\nabla d(x) = \alpha(x) \Pi(x) x, \tag{12}$$

where $\alpha(x)$ is a positive scalar function, and $\Pi(x)$ is the *Scaling Matrix*, which is always chosen to be *diagonal*. An important distinction amongst the diversity measures from an algorithmic point of view is whether their scaling matrix is positive definite or not. For diversity measures with positive definite scaling matrices, we have been able to develop simpler convergent algorithms.

4.2. A Generalized Affine Scaling Algorithm

The affine scaling methodology developed in [9] is readily adapted to address the minimization of the more general diversity measures developed in [6]. The first order necessary condition for a solution to the concave constrained minimization problem,

$$\min_x d(x) \quad \text{subject to} \quad Ax = b\,, \tag{13}$$

naturally suggests an iterative algorithm of the form [9, 6]

$$x_{k+1} = \Pi^{-1}(x_k)A^T(A\Pi^{-1}(x_k)A^T)^{-1}b\,. \tag{14}$$

This algorithm has desirable properties when the scaling matrix $\Pi(x)$ is positive definite. As shown in [9, 6], when $\Pi(x)$ is positive definite it can be used to naturally define an Affine Scaling Transformation (AST) matrix, $W(x) = \Pi^{-\frac{1}{2}}(x) > 0$, and thereby establish a strong connection to affine scaling methods used in optimization theory [5]. Hence the use of the terminology "Affine Scaling" in connection with the algorithms developed here and in [9].

Following the AST methodology [5], the scaled quantities q_k and A_{k+1} are defined by

$$W_{k+1} = W(x_k)\,, \quad x_k = W_{k+1}q_k\,, \quad A_{k+1} = AW_{k+1}\,,$$

assuming we have at hand a current estimated feasible solution, x_k (equivalently, q_k) to the problem (13). We can then recast the optimization problem (13) in terms of the scaled variable $q = W_{k+1}^{-1}x$,

$$\min_q d_{k+1}(q) \triangleq d(W_{k+1}q) \quad \text{subject to} \quad A_{k+1}q = b\,.$$

We then obtain an *updated* feasible estimate, q_{k+1}, by projecting the gradient of $d_{k+1}(q_k)$ onto the nullspace of A_{k+1} and moving in this direction an amount proportional to a stepsize given by $1/\alpha(x_k)$ [9, 6]. This yields the algorithm

$$q_{k+1} = A_{k+1}^{+}b\,, \quad x_{k+1} = W_{k+1}q_{k+1}\,,$$

with A_{k+1}^{+} the Moore-Penrose pseudoinverse of A, which is equivalent to (14).

It is common in the affine scaling approach to use the specific AST $W(x)$ given by

$$W(x) = \text{diag}\left(|x[i]|\right)\,,$$

which corresponds to defining $W(x)$ in terms of $\Pi(x) = \Pi_p(x)$ for $p = 0$. In contrast, the algorithm (14) corresponds to a "natural" choice of $W(x)$ dictated by the particular choice of the sparsity measure $d(x)$.

Convergence of the algorithm can be shown for specific classes of diversity measures, $d(x)$, using the general convergence theorems of Zangwill, and their variants [14, 9]. The strongest results hold when a sign and permutation invariant concave diversity measure $d(x)$ has a positive definite scaling matrix, $\Pi(x) > 0$ for all $x \in \mathrm{R}^n$. This condition does not appear to be overly restrictive and it admits the large class of $\mathcal{S}$-Functions described in Section 3.1. This class satisfies the conditions of Theorem 12 and also contains the p-norm-like, $p \leq 1$, concave sparsity measures.

Theorem 12 [6] *Let $d(x)$ be a sign and permutation invariant function that is strictly concave on the positive orthant $\mathcal{Q}_1$ and for which $\Pi(x) > 0$ for all $x \in \mathrm{R}^n$. Assume that the set $\{x|d(x) \leq d(x_0)\}$ is compact for all x_0. Let x_k be generated by the iteration (14) starting with x_0 feasible, $Ax_0 = b$. Then for all $\bar{x}_{k+1} \neq \bar{x}_k$, we have $d(x_{k+1}) < d(x_k)$ and the algorithm converges to a local minimum $d(x^*)$, $x_k \to x^*$, where x^* is a boundary point of $\mathcal{Q}_l \cap LV(A, b)$ for some orthant $\mathcal{Q}_l$.*

Other functions can be proved to be minimized by the algorithm (14), with convergence generally being shown on a case-by-case basis. For example, it is proved in [9] that the 2-normalized Shannon entropy–based algorithm is convergent.

As discussed above, the 2-normalized Shannon entropy diversity measure corresponds to the 2-normalized Renyi entropy for $p = 1$, which is not concave or Schur-concave, and therefore is not expected to result in a minimum associated with complete sparsity; a fact demonstrated in simulation [9]. Lack of concavity requires a convergence proof via different means than the use of (4).

The requirement of the invertibility of $\Pi(x)$ in (14) appears to give good reason to prefer the measures provided by the class of (unnormalized) $\mathcal{S}$-functions over the 1-norm and 2-normalized scale invariant $\mathcal{S}$-functions (which effectively include the normalized Renyi entropies). In particular, the tractable form of the scaling matrices for $\mathcal{S}$-functions allows them to be readily inverted [6].

5. REFERENCES

[1] J. Adler, B. Rao, & K. Kreutz-Delgado, "Comparison of Basis Selection Methods," *Proc. Asilomar Conf. Signals, Sys., Comp.*, 1996.

[2] E. Beckenbach & R. Bellman, *Inequalities*, Spr.-Ver., 1971.

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

[4] D. Donoho, "On Minimum Entropy Segmentation" in C.K. Chui, L. Montefusco, and L. Puccio, ed.s, *Wavelets: Theory, Algorithms, and Applications*, pp. 233-269, *AP*, 1994.

[5] S.-C. Fang & S. Puthenpura, *Linear Optimization and Extensions: Theory and Algorithms*, Prentice Hall, 1993.

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

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

[8] G. Patil & C. Taillie, "Diversity as a Concept and its Measurement," *J. Amer. Stat. Assoc.*, 77(379):548–67, 1982.

[9] B.D. Rao & K. Kreutz-Delgado, "An Affine Scaling Methodology for Best Basis Selection," UCSD CIE Report, 1997. Submitted to *IEEE Trans. Signal Proc.*

[10] B.D. Rao, "Signal Processing with the Sparseness Constraint," *ICASSP 98*.

[11] A. Renyi, *Probability Theory*, North-Holland, 1970.

[12] R.T. Rockafellar, *Convex Analysis*, Princeton, 1970.

[13] M. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A.K. Peters, Wellesley, MA, 1994.

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