A Globally Convergent Algorithm for MAP Estimation in the Linear Model with Non-Gaussian Priors

J. A. Palmer
ECE Department
Univ. of California San Diego
La Jolla, CA 92093
japalmer@ucsd.edu

K. Kreutz-Delgado *
ECE Department
Univ. of California San Diego
La Jolla, CA 92093
kreutz@ece.ucsd.edu

Abstract

We develop a framework for analyzing non-gaussian densities in terms of the curvature of the density function itself, rather than moments of the random variable. The framework suggests a new criterion for sub- and super-gaussianity of densities that is seen to be of wider range of application than the commonly used kurtosis criterion. We show that the notion of relative curvature introduced can be seen as a generalization of the notion of convexity, where classical convexity of a function is seen as a relationship between the function and a linear model. We use the curvature framework to derive an inequality that holds for all functions that are super-gaussian in the sense of the proposed criterion. This inequality allows proof of global convergence of a certain re-weighted minimum norm algorithm by providing a weighting matrix that yields descent without line search. The algorithm is equivalent to the FOCUSS algorithm of [1, 2] in the case of independent Generalized Gaussian densities in the linear model.

1 Introduction

The present research was inspired by the problem of £nding sparse solutions to a linear inverse problem. We approach the problem using the Bayesian framework of [3]. Let $y \in \mathbb{R}^m$ be an observed instance of the random vector Y, which is modelled according to the linear model,

$$Y = AX + \nu$$

where $A \in \mathbb{R}^{m \times n}$ is a deterministic matrix, X is an n-dimensional random source vector, and ν is an m-dimensional random noise vector, with all component random variables mutually independent. The MAP estimation problem is to £nd $\hat{x}$, de£ned by,

$$\hat{x} = \arg\max_{x \in \mathbb{R}^n} p_{X|Y}(x|y) \tag{1.1}$$

*This research was partially supported by NSF Grant No. CCR-9902961

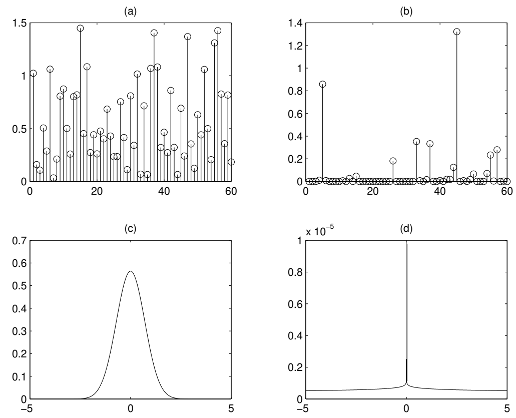

Figure 1: (a) shows a typical vector generated by the (non-sparse) Gaussian density shown in (c), and (b) shows a typical vector generated by the (sparse) generalized Gaussian density (with shape parameter 0.1) shown in (d)

The sources are considered to be sparse if the source component densities are sharply peaked around zero and "heavy-tailed". One can think of a sparse random variable as having two possible states: active and inactive. When inactive, the random variable takes a value close to zero, and when active, it takes a value far enough from zero to be easily distinguishable from the inactive zero state [4]. The "sparsity" of the random variable is a result of the fact that it has a higher probability of being in the inactive state. For example, neurons are often modelled as sparse random variables [5], and similarly the random variables corresponding to the components of wavelet dictionaries [6], and to the independent components learned in ICA [5] are also well-known to be sparse.

The "sparseness" of the random variable is related to the curvature of the density. With the requirement that the density be unimodal, in order to have a signi£cant mass around zero (large probability of being inactive) and a large expected magnitude (taking on a relatively large value when active), the density must peak sharply at zero, around the inactive range, and ¤atten out, descending slowly over the active range.

The commonly used measure of this curvature, and thus of the sparsity of the random variable, is the kurtosis, which, for a zero mean random variable X, is the difference between the fourth moment of X and the fourth moment of a Gaussian random variable with the same variance as X, i.e. $E\,X^4 - 3\,(E\,X^2)^2$. The kurtosis is positive for super-gaussian densities and negative for sub-gaussian densities. While the kurtosis measure is consistent with the notion of sparsity given above, it relies on the existence of even moments, and is thus not always £nite. As sparse densities tend to have non-integrable moments, the kurtosis measure can fail in comparing very sparse or super-gaussian densities. For example, the inverse tangent can be used as a non-linearity in source separation networks [5]. Such a network performs the equivalent of MAP estimation of the source assuming a Cauchy prior. It is clear that the Cauchy density has the characteristics of a super-gaussian density, but we are unable to order this density with respect to other super-gaussian densities using the kurtosis criterion. Even if the moments of the density are £nite, it may not be possible to evaluate the integral in closed form.

In this paper we propose a differential criterion for super-gaussianity of a density that is also "centered" in a sense at the Gaussian density. The differential criterion arises from a consideration of the relative curvature of two functions, and may be seen as a generalization of the notion of convexity, where convexity is seen as comparing a function to the linear model of the function at one or two points. We propose to consider densities sub- or super-gaussian as the negative log densities are convex or concave relative to the negative log Gaussian, i.e. relative to the quadratic function. A happy result of this de£nition of super-gaussianity is that it yields a natural re-weighted minimum norm algorithm that is globally convergent without line search for all super-gaussian densities thus de£ned. Furthermore, we have the result that the Fenchel-Legendre conjugate of (the log of) a super-gaussian density is sub-gaussian, and vice-versa. Thus the dual problem of a linear MAP estimation problem with sub-gaussian prior is a super-gaussian estimation problem, so the proposed algorithm also applies to sub-gaussian densities.

2 Function Curvature

Let f be a continuously differentiable increasing function over a possibly unbounded interval (a, b). We wish to de£ne a measure of the curvature of f at a point in (a, b). We proceed £rst in an operational manner. Given two functions, f and h, we wish to say when f is has "greater curvature" than h at $x_0 \in (a, b)$. De£ne the tangent function $T_{f,h}$ of f applied to h at the point $x_0 \in (a, b)$ to be the unique af£ne transformation of $f(x)$, $T_{f,h}(x_0, x) = \alpha(x_0) f(x) + \beta(x_0)$, that gives,

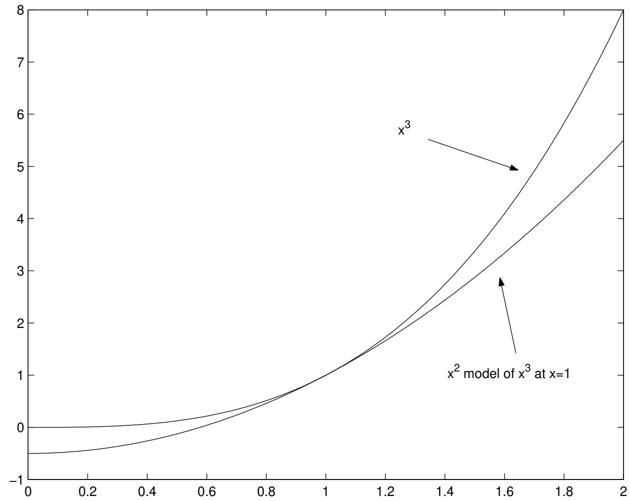

Figure 2: The function $T_{f,h}$ is de£ned at x_0 by $T_{f,h}(x_0, x) = \alpha(x_0) f(x) + \beta(x_0)$ such that it is tangent to $h(x)$ at x_0. h has more curvature than f if $T_{f,h}(x_0, x) \le h(x)$ for $x \in (a, b)$.

$$T_{f,h}(x_0, x_0) = h(x_0) \quad \text{and} \quad T'_{f,h}(x_0, x_0) = h'(x_0) \tag{2.1}$$

where h' is the derivative of h. That is, $T_{f,h}$ is af£nely related to f so that it has function value and slope equal to the value and slope of h at x_0, i.e. $T_{f,h}$ and h are tangent at x_0. Working with the tangent model $T_{f,h}$ instead of f in a sense removes the in¤uence of £rst order information. As a provisional de£nition, if f and h are convex, we shall take the curvature of f to be greater than that of h at $x_0 \in (a, b)$ if the tangent function of f applied to h at x_0 lies uniformly above h on (a, b). That is,

$$T_{f,h}(x_0, x) \ge h(x) \quad \forall\, x \in (a, b) \tag{2.2}$$

See Figure 2. Similarly, for f and h concave, f is de£ned to have greater curvature than h at x_0 if the inequality (2.2) is reversed. For convex f and h, we say that f has greater curvature than h on (a, b) if,

$$T_{f,h}(x, y) \ge h(y) \quad \forall\, x, y \in (a, b) \tag{2.3}$$

and again the inequality is reversed in the case of f and h concave. For £xed x_0, the univariate function $T_{f,h}(x_0, x)$ is an af£ne function of $f(x)$ de£ned by $\alpha(x_0) f(x) + \beta(x_0)$. It can be considered a £rst order model of the function h at x_0 using f. The parameters $\alpha(x_0)$ and $\beta(x_0)$, chosen to satisfy (2.1), are easily seen to be,

$$\alpha(x_0) = \frac{h'(x_0)}{f'(x_0)} \qquad \beta(x_0) = h(x_0) - \frac{h'(x_0)}{f'(x_0)} f(x_0)$$

Substituting these into (2.2), we have that f has greater curvature than h on (a, b) if,

$$f(x) - f(x_0) \geq \frac{f'(x_0)}{h'(x_0)} \big(h(x) - h(x_0)\big) \quad \forall\, x \in (a, b) \tag{2.4}$$

It may be veri£ed that this is equivalent to de£ning h to have "less curvature" than f at x_0 if the af£ne transformation of $h(x)$, $T_{h,f}(x_0, x)$, lies entirely below $f(x)$ on (a, b). The equation (2.4) is similar to the differential de£nition of convexity, and indeed reduces to this de£nition when the function h is af£ne. In fact, (2.4) is the differential de£nition of convexity for the composite function $f \circ h^{-1}$.

3 Relative Convexity and Square Convexity

Based on the considerations in the previous section, we de£ne relative convexity as follows.

De£nition (Relative Convexity). For f and h continuous and increasing on the interval (a, b), f is *convex relative to* h on (a, b) if $f \circ h^{-1}$ is convex on $\big(h(a), h(b)\big)$.

Note that this de£nition does not assume differentiability of f and h, only invertibility on (a, b). We can show using basic properties of continuous, increasing, and de£nite functions (not necessarily differentiable, but convex or concave on an interval), that relative convexity, considered as a relation on the set of functions continuous and increasing on (a, b), induces a partial ordering on that set [7]. Thus we can write $f \succeq h$ or $f \preceq h$ on (a, b) when f has greater or lesser curvature than h respectively. Also f is convex relative to h if and only if h is concave relative to f.

Since relative convexity is de£ned in terms of the convexity of a composite function, we can apply the theory of differentiable convex functions to the composite function to derive the following.

Lemma 1 (Relative convexity for differentiable functions). *If f and h are increasing and differentiable on (a, b), then $f \succeq h$ on (a, b) if and only if,*

$$f(y) - f(x) \geq \frac{f'(x)}{h'(x)} \big(h(y) - h(x)\big) \quad \forall\, x, y \in (a, b) \tag{3.1}$$

If f and h are increasing and twice differentiable on (a, b), then $f \succeq h$ on (a, b) if and only if,

$$\frac{f''(x)}{f'(x)} \geq \frac{h''(x)}{h'(x)} \quad \forall\, x \in (a, b) \tag{3.2}$$

Abstracting from (3.2), the operator $K : f \rightarrow \frac{f''}{f'}$, which is equivalent to the operator $D \log D$, can be seen as a measure of instantaneous order that is invariant to af£ne transformation of $f(x)$ [7]. In fact, $\frac{f''(x)}{f'(x)}$ is the exponent γ in an exponential model of f, $\alpha \exp(\gamma x) + \beta$, de£ned so that it agrees with f in function value, and £rst and second derivatives.

We can generalize the de£nition of relative convexity to derive useful inequalities for functions de£ned on $\mathbb{R}^n$.

De£nition (Relative Convexity in $\mathbb{R}^n$). Let $C \subset \mathbb{R}^n$ be convex, and let $f : C \rightarrow \mathbb{R}$, $g : C \rightarrow C$. We say that f is *convex relative to* g on C if $f \circ g^{-1}$ is convex on $g(C)$.

We now derive the differential implications of the multidimensional de£nition. In the following $J(x)$ and $H(x)$ will be used for Jacobians and Hessians, with J_g and H_f referring for example to the Jacobian of g and the Hessian of f. The gradient of f will be denoted ∇f.

Lemma 2 (Relative convexity for differentiable functionals). *Let $C \subset \mathbb{R}^n$, and let $f : C \rightarrow \mathbb{R}$, $g : C \rightarrow C$, with ∇f and J_g continuous, and ∇f invertible on C. Then f is convex relative to g on C if and only if*

$$f(y) - f(x) \geq \nabla f(x)^T J_g(x)^{-1} \big(g(y) - g(x)\big) \quad \forall\, x, y \in C$$

If H_f and H_{g_k} are invertible on C, where g_k is the kth component of g, then f is convex relative to g if and only if

$$H_f(x) \geq \sum_{k=1}^{n} H_{g_k}(x) \big[J_g(x)^{-T} \nabla f(x)\big]_k \quad \forall\, x \in C$$

where the notation $A \geq B$ is used for $A - B$ positive semide£nite.

In the particular case of $g(x) = x^2$, the component-wise squaring operation, we have $J_g(x) = \mathrm{diag}(2x)$. De£ne $W = \mathrm{diag}(x)$. Then for f concave relative to x^2 we have,

$$f(y) - f(x) \leq \frac{1}{2} \nabla f(x)^T W^{-1} \big(y^2 - x^2\big) \quad \forall\, x, y \in C \tag{3.3}$$

where diagonal components of f may be in£nite if the corresponding component of x is zero. We shall call functions that are concave relative to x^2, "square-concave", and those that are convex relative to x^2, "square-convex". For f twice differentiable, we have f square-concave if,

$$\frac{y^T H_f(x) y}{y^T y} \leq \big\| W^{-1} \nabla f(x) \big\|_1 \quad \forall\, x, y \in C \tag{3.4}$$

that is, if the 2-norm of the matrix H_f projected onto C is less than the 1-norm of $W^{-1} \nabla f(x)$ for all $x \in C$.

4 Algorithm for Linear Estimation with Super-Gaussian Densities

Let $f:\mathbb{R}^n \to \mathbb{R}$ be separable, so that $f(x) = \sum_i f_i(x_i)$, and let each component function $f_i : \mathbb{R} \to \mathbb{R}$ be sign-invariant (symmetric about zero) and increasing with the magnitude of its argument, i.e. $f(x) = f(|x|)$, and $|y_i| > |x_i|$ implies $f_i(y_i) > f_i(x_i)$. The increasing property implies that x and $\nabla f(x)$ are in the same orthant for all x, so that that $W\nabla f(x) \geq 0$, where $W \equiv \mathrm{diag}(x)$.

Consider the problem $\min_{x\in C} f(x)$, where C is a convex set, e.g. the linear variety de£ned by $Ax = b$, and f is square-concave on C. We can use the inequality (3.3) to de£ne a descent algorithm as follows. At iteration k, for any x_{k+1}, using the separability of f we have,

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \frac{1}{2}\nabla f(x)^T W_k^{-1}\left(x_{k+1}^2 - x_k^2\right) \\ &\equiv \frac{1}{2}\, x_{k+1}^T \Pi_k\, x_{k+1} - \frac{1}{2}\, x_k \Pi_k\, x_k \end{aligned}$$

where $W_k \equiv \mathrm{diag}(x_k)$, and $\Pi_k \equiv \mathrm{diag}(W_k^{-1}\nabla f(x)) \geq 0$. Then if we take for x_{k+1},

$$x_{k+1} \leftarrow \arg\min_{x\in C}\, x^T \Pi_k\, x \tag{4.1}$$

we can guarantee that $f(x_{k+1}) \leq f(x_k)$. When C is the linear variety de£ned by $Ax = b$, the minimization can be carried out by solving,

$$\begin{bmatrix} \Pi_k & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix} \tag{4.2}$$

This proves descent of the algorithm. De£ning the solution set to be the set of £xed points of the algorithm, we can show global convergence on a large class of convex sets [7] using one of Zangwill's global convergence theorems [8]. When C is a polytope, including the case where C is a linear variety, the set of £xed points is bounded, and we have the following.

Theorem 1 (Global convergence on polytopes [7]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be separable, sign-invariant, and increasing over each orthant, and let f be square-concave on the convex polytope C. Then the algorithm* (4.1) *converges to a local minimum of f on C from any point in $\mathbb{R}^n$ except possibly on a set of measure zero.*

We can apply this theory to the linear MAP estimation problem (1.1) by considering the curvature of the negative log prior densities.

$$\begin{aligned} \hat{x} &= \arg\min_x -\log p_X(x) - \log p_{Y|X}(y|x) \\ &= \arg\min_x \sum_{i=1}^n f_i(x_i) + \sum_{j=1}^m d_j(y_j - \bar{a}_j^T x) \end{aligned}$$

where the component prior densities are assumed to be mutually independent, $\bar{a}_j$ is the jth row of A, $f_i(x_i) \equiv -\log p_{X_i}(x_i)$ and $d_j(y_j - \bar{a}^T x) \equiv -\log p_{Y_j|X}(y_j|x)$.

Like the kurtosis measure, the relative convexity measure coincides with the notion of sub- and super-gaussianity if we take the density p_X to be super-gaussian when $-\log p_X$ is square-concave (or p_X negative log square-concave). In addition, densities that are commonly taken to be super-gaussian but do not have moments, like the Cauchy, are amenable to the relative convexity measure. In the following we assume that the source components f_i and the noise components d_j are square-concave (which includes the cases of Gaussian noise and the L_1-norm error). De£ning $e \equiv y - Ax$, the problem can be written,

$$\hat{x} = \arg\min_{x,e} \sum_{i=1}^n f_i(x_i) + \sum_{j=1}^m d_j(e_j) \quad s.t. \quad Ax + e = y$$

Then de£ning $C \equiv [\,A \;\; I\,]$ and $z \equiv \left[\,x^T \;\; e^T\,\right]^T$, we have

$$\hat{z} = \arg\min_z \sum_{i=1}^{n+m} \tilde{f}_i(z_i) \quad s.t. \quad Cz = y$$

where we de£ne $\tilde{f}_i$ to range over the f_i and d_j functions, each of which is assumed square concave.

Now we can apply the algorithm to this problem to £nd the estimate $\hat{x}$. If z is updated by,

$$z_{k+1} \leftarrow \arg\min_z z^T \Pi_k\, z \quad s.t.\; Cz = y$$

we can guarantee descent of $\tilde{f}$. This minimization can be carried out by solving,

$$\begin{bmatrix} \Pi_k(x_k) & 0 & A^T \\ 0 & \Pi_k(e_k) & I \\ A & I & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} \\ e \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ y \end{bmatrix}$$

where $\Pi_k(x_k) \equiv \mathrm{diag}(W_k^{+}\nabla f(x_k))$ and $\Pi_k(e_k) \equiv \mathrm{diag}(V_k^{+}\nabla d(e_k))$, and W_k^+ and V_k^+ are pseudoinverses obtained by inverting the non-zero components of $\mathrm{diag}(x_k)$ and $\mathrm{diag}(e_k)$ respectively. We can write x_{k+1} in closed form as,

$$x_{k+1} = \Pi_k^+(x_k)A^T\left(A\,\Pi_k^+(x_k)A^T + \Pi_k^+(e_k)\right)^{-1} y$$

It is unnecessary to solve for e_{k+1} as it is constrained to be $y - Ax_{k+1}$. The recommended method of computing x_{k+1} is to £rst solve,

$$\left(A\,\Pi_k^+(x_k)A^T + \Pi_k^+(e_k)\right)\lambda = y$$

and then let $x_{k+1} = \Pi_k^+(x_k)A^T\lambda$.

5 Conjugate Symmetry with respect to Quadratic and Logarithmic Curvatures

As further evidence of the naturalness of the de£nition of curvature in terms of relative convexity, we give a relationship between square-convexity, convexity relative to $\log$, and Fenchel-Legendre conjugacy of one dimensional functions. The relationship is also easily extended to separable functionals. We assume in the following that $f:\mathbb{R}\rightarrow\mathbb{R}$ is twice differentiable and de£nite on (a,b). De£niteness of f on (a,b), i.e. strict positivity or negativity of f'' on (a,b) ensures that f' is invertible on (a,b) as it is then strictly increasing or strictly decreasing.

The Fenchel conjugate of a convex function f, denoted f^*, is de£ned by [9],

$$f^*(\phi) = \sup_x \; \phi\, x - f(x)$$

If f is concave, then the Fenchel conjugate is de£ned by, $f^*(\phi) = \inf_x \; \phi\, x - f(x)$. Since we assume that f is smooth and de£nite on (a,b), the Fenchel conjugate is the same as the Legendre transform [9]. For the derivatives, we have,

$$f^{*\prime}(\phi) = f'^{-1}(\phi) = x(\phi)$$
$$f^{*\prime\prime}(\phi) = f'^{-1\prime}(\phi) = \frac{1}{f''\big(f'^{-1}(\phi)\big)} = \frac{1}{f''\big(x(\phi)\big)}$$

We now have the following theorem.

Theorem 2 (Fenchel-Legendre conjugate symmetry). *Let f be increasing and de£nite on $\mathcal{D}$, with Fenchel conjugate f^* de£ned on $\mathcal{D}^* = \{\phi : f'(x) = \phi,\ x \in \mathcal{D}\}$. Then f is square-convex on $\mathcal{D}$ if and only if f^* is square-concave on $\mathcal{D}^*$, and f is concave relative to* $\log$ *on $\mathcal{D}$ if and only if f^* is convex relative to* $\log$ *on $\mathcal{D}^*$.*

Proof. We have f square-convex if and only if $f''(x)/f'(x) \geq 1/x$. Since under our assumptions f^* is twice differentiable on $\mathcal{D}^*$, we have f square-convex if and only if,

$$\frac{f^{*\prime\prime}(\phi)}{f^{*\prime}(\phi)} = \frac{1}{x(\phi) f''(x(\phi))} \leq \frac{1}{f'(x(\phi))} = \frac{1}{\phi}$$

that is, if and only if f^* is square-concave on $\mathcal{D}^*$.

Similarly, f is concave relative to $\log$ if and only if $f''(x)/f'(x) \leq -1/x$. But this holds if and only if,

$$\frac{f^{*\prime\prime}(\phi)}{f^{*\prime}(\phi)} = \frac{1}{x(\phi) f''(x(\phi))} \geq -\frac{1}{f'(x(\phi))} = -\frac{1}{\phi}$$

that is, if and only if f^* is convex relative to log. □

Since the Fenchel-Legendre conjugate is used in the dual of the MAP optimization problem, we can turn a sub-gaussian problem into a super-gaussian problem, and then apply the globally convergent algorithm given here. This theoretical global convergence is also shown to be useful in deriving convergent dictionary learning and ICA algorithms in [10]. The algorithm given is super-linear for log concave densities, as can be seen by generalizing the convergence rate results in [1], but it is linear for $\log$ square-concave densities that are $\log$ convex, wither higher asymptotic constant the closer f is to quadratic.

If we admit line search, we can also use Theorem 2 to achieve stable quadratic convergence by applying Newton's method to the dual. This is the idea proposed in [11] for discrete L_p optimization with $1 < p < 2$. Newton's method is unstable when applied to the primal problem as the second derivative of super-gaussian functions is generally unbounded at zero. The result given here regarding the conjugate operation and "re¤ection" about quadratic curvature, yields insight into the effect on curvature of the conjugacy operation, and may be used to derive a general alternative theorem concerning the boundedness of the second derivative of either the primal or the dual.

References

[1] B. D. Rao and I. F. Gorodnitsky. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Trans. Signal Processing*, 45:600–616, 1997.

[2] B. D. Rao and K. Kreutz-Delgado. An af£ne scaling methodology for best basis selection. *IEEE Trans. Signal Processing*, 47:187–200, 1999.

[3] B. D. Rao and K. Kreutz-Delgado. Basis selection in the presence of noise. In *Proceedings of the 1998 Asilomar Conference*. IEEE, 1998.

[4] S. A. Kassam. *Signal Detection in Non-Gaussian Noise*. Springer-Verlag, New York, 1988.

[5] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal of Scienti£c Computation*, 20(1):33–61, 1998.

[7] J. A. Palmer. Function curvature, relative convexity, and conjugate curvature. Technical report, ECE Dept., UCSD, 2002.

[8] W. I. Zangwill. *Nonlinear Programming: A Uni£ed Approach*. Prentice-Hall, 1969.

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

[10] J. A. Palmer and K. Kreutz-Delgado. A general framework for dictionary learning and ICA. Technical report, ECE Dept., UCSD, 2002.

[11] J. Fischer. An algrorithm for discrete linear L_p approximation. *Numerische Mathematik*, 38(1):129–139, 1981.