A GENERAL FRAMEWORK FOR COMPONENT ESTIMATION

J. A. Palmer and K. Kreutz-Delgado

Dept. of Electrical and Computer Engineering,
Univ. of California San Diego, La Jolla, CA 92093

ABSTRACT

Component estimation arises in Independent Component Analysis (ICA), Blind Source Separation (BSS), wavelet analysis and signal denoising [1], image reconstruction [2, 3], Factor Analysis [4], and sparse coding [5, 6]. In theoretical and algorithmic developments, an important distinction is commonly made between sub- and super-gaussian densities, super-gaussian densities being characterized as having high kurtosis, or having a sharp peak and heavy tails. In this paper we present a generalized convexity framework similar to a classical concept of E. F. Beckenbach [7], which we refer to as relative convexity. Based on a partial ordering induced by relative convexity, we derive a new measure of function curvature and a new criterion for super-gaussianity that is both simpler and of wider application than the kurtosis criterion. The relative convexity framework also provides an inequality that can be used to derive stable and effective descent algorithms for estimation of the parameters in the Bayesian linear model when sub- or super-gaussian priors are used. Apparently almost all common symmetric densities are comparable in this ordering to Gaussian, and thus are either sub- or super-gaussian, despite the fact that the measure is instantaneous, in contrast to moment-based measures. We present several algorithms for component estimation that are shown to be descent algorithms based on the relative convexity inequality arising from the assumption of super-gaussian priors. We also show an interesting relationship between the curvature of a convex or concave function and the curvature of its Fenchel-Legendre conjugate, which results in an elegant duality relationship between estimation with sub- and super-gaussian densities.

1. INTRODUCTION

Our research was inspired by the problem of learning data representations based on a linear generative model [8, 5, 9, 6, 10, 11, 12, 13]. Given observations $\mathbf{Y} = [y_1 \ldots y_N]$, the problem is to estimate the parameters $A \in \mathbb{R}^{m \times n}$ and $\mathbf{X} = [x_1 \ldots x_N]$ in the Bayesian linear model,

$$y_k = Ax_k + \nu_k, \quad k = 1, \ldots, N \tag{1.1}$$

assuming that the sources x_k are independent. The low noise limit is equivalent to the case in which the noise random variables ν_k are not present. Since Field [8], much consideration has been given to representations that assume sparse and distributed sources, i.e. many source components with relatively few of the components having significant magnitude, or "active", at any given time. One way to ensure a sparse representation is to take A to be "overcomplete", or have more columns than rows. However, sparse coding can also be carried out when the matrix A is not overcomplete, for example when the data is high dimensional but occupies a relatively low dimensional manifold [14].

In the complete and undercomplete cases $A \in \mathbb{R}^{m \times n}$, $m \geq n$, various information theoretic criteria can be used to obtain algorithms for estimating A [15, 16, 17, 18]. Many of these methods are known to be equivalent to gradient based algorithms for Maximum Likelihood (ML) estimation assuming certain sparse, or super-gaussian priors. In the overcomplete case, $A \in \mathbb{R}^{m \times n}$, $m < n$, it is common to start from the statistical framework.

Two main statistical approaches have been used in the estimation of overcomplete A: an ML approach estimating A by (approximately) marginalizing over $\mathbf{X}$, and a joint maximum *à posteriori* (MAP) estimation approach estimating both A and $\mathbf{X}$. Lewicki and Olshausen [9] and Lewicki and Sejnowski [6] propose an approximate ML framework. The likelihood is marginalized over the possible generating sources,

$$p(\mathbf{Y}|A) = \int p(\mathbf{Y}, \mathbf{X}|A)\, d\mathbf{X}$$

and the resulting integral is approximated by a Gaussian density. Various other approximations are made to obtain a generalization of the Infomax algorithm of Bell and Sejnowski [16]. Girolami [13] also uses the ML approach, but employs a variational approximation similar to those used by Jaakkola [19]. An algorithm is derived using a variational bound on the Laplacian (sparse) prior that makes the marginalization tractable. The EM algorithm is used to optimize the variational likelihood over the augmented set of parameters. As the variational parameters converge, the variational approximation approaches Laplacian. This algorithm apparently converges to a local minimum of $p(\mathbf{Y}|A)$ itself rather than an approximation as in [6]. It is also globally convergent via the EM algorithm. As usual, however, the algorithm is subject to convergence to local optima.

Another approach is to find the joint estimate of A and $\mathbf{X}$. Here we maximize,

$$p(A, \mathbf{X}|\mathbf{Y}) = p(\mathbf{Y}|A, \mathbf{X})\, p(A, \mathbf{X})$$

For the optimization to be well defined, a constraint must be put on either $\mathbf{X}$ or A, which amounts to determining

This research was partially supported by NSF Grant No. CCR-9902961. Authors can be contacted at {japalmer,kreutz}@ucsd.edu.

$p(A, \mathbf{X})$ [15, 10, 12]. Olshausen and Field [5] used a cost function equivalent to that implied by the MAP framework. The MAP framework is used explicitly in Hyvärinen [10] for complete and undercomplete A, and in Kreutz-Delgado and Rao [12] for overcomplete A. [1] Algorithms are also found in [11]. The usual idea is to view the log likelihood as consisting of an error term and a sparsity term, and alternately adapt $\mathbf{X}$ and A, with $\mathbf{X}$ adapting to increase the sparsity of the representation, and A adapting to maintain fidelity of the representation.

In this paper we employ the MAP framework to estimate A and $\mathbf{X}$ given $\mathbf{Y}$ in a novel way. Rather than adapting A to reduce an error term only, thus reducing the sparsity of the representation only indirectly, we derive a general procedure for adapting A to reduce the source prior term directly in the noiseless case, and show that the noiseless analysis can be applied to the noisy case as well by a change of variables. The noiseless algorithms can be seen as new generalizations of the Infomax algorithm [16] to the overcomplete case, which differ from that given in Lewicki and Sejnowski [6]. The noisy algorithms are similar to the approximate algorithms found in [10] for the complete and undercomplete cases. Our analysis however does not require A to be invertible, and we derive a descent algorithm for the original cost function, not an approximation. The generality of the approach allows application to the undercomplete dictionary case as well, which may be useful when the intrinsic dimensionality of the data is less than the dimensionality of the observed vectors [14]. For example, an image compression scheme might code 12 pixel $\times$ 12 pixel blocks, but the image blocks that are coded may have a simple structure that is representable in fewer than the 144 basis vectors that result from using even a complete representation, much less an overcomplete representation.

The organization of the paper is as follows. In section 2 we discuss the results of an investigation into the notion of sub- and super-gaussianity. We derive a criterion for sub- and super-gaussianity that is seemingly more natural and of wider application than the commonly used kurtosis criterion. The criterion is based on a generalization of the notion of convexity which we call relative convexity. In section 3, we show that the optimization problem associated with the MAP estimation framework can always be formulated as an equality constrained nonlinear optimization problem, and we apply the relative convexity theory to derive a descent algorithm for estimating super-gaussian components or sources with a given A. The proposed criterion for super-gaussianity applies to all of the densities commonly taken to be super-gaussian, including the Cauchy density. In section 4, we use the relative convexity framework to derive algorithms for component estimation. We first derive a globally convergent algorithm for A assumed to have unit Frobenius norm. This algorithm is suggestive of a sort of Lagrangian fixed point algorithm. We formulate the Lagrangian and optimality conditions for A and $\mathbf{X}$, and propose a general procedure for deriving fixed point algorithms for the estimation of A and $\mathbf{X}$ corresponding to different priors on A. We give algorithms for the particular cases of $p(A)$ uniform over unit Frobenius norm matrices $\|A\|_F = 1$, and $p(A)$ uniform over unit column norm matrices $A = [a_1 \cdots a_n]$, $\|a_i\| = 1$, $i = 1 \ldots n$. Finally we state a duality result for relative convexity that allows the analysis to be applied to sub-gaussian estimation as well.

[1] It is commonly assumed that the MAP approach is a simplified version of the more appropriate ML framework. Theoretical arguments aside, in our experiments the ML algorithm of [13] seemed to be subject to more local optima and degenerate solutions than the MAP algorithms.

2. SUPER-GAUSSIANITY, RELATIVE CONVEXITY AND SQUARE-CONCAVITY

2.1. The measure of super-gaussianity

Along with the sparsity criterion, Field [8] advocates kurtosis as a measure of sparsity. The kurtosis of a zero mean random variable X can be defined as the difference between the fourth moment of X and the fourth moment of a Gaussian random variable of equal variance, or $E(X^4) - 3E(X^2)^2$. If a density has positive kurtosis, then it is likely to be more peaked about the mean, and have "heavier tails" than the Gaussian density. Such a density is commonly called super-gaussian, while a density with negative kurtosis is called sub-gaussian. The kurtosis criterion is not without controversy (see [20, §6] and references therein) but in general it coincides with the intuitive notion of sparsity of a random variable, that it be more likely to be either zero (inactive) or relatively large in magnitude (active) with little probability of "in between" values. A key feature of the kurtosis is its ordering of densities with respect to the Gaussian density, with sub-gaussian densities on one side and super-gaussian densities on the other.

In [21] an operational approach is taken to the partial ordering of densities with respect to Gaussian. The concept of "over-gaussianity" is defined as a density's having a tail that is asymptotically heavier than than the Gaussian tail, with sub-gaussianity defined similarly. A theorem is given that for a unimodal density having two points of intersection with the normalized Gaussian density, the density is over-gaussian if and only if the density has positive kurtosis. Comparing the (asymptotic) heaviness of the tails is a natural way to compare densities in this context, and provides a relatively simple way to compare densities with Gaussian. We could similarly compare the order of the negative log density to the order of x^2. The comparison of asymptotic order however does not address the issue of peakedness. A similar theorem is given in Finucan [22] which assumes four density crossings (both sharper peak and heavier tails) rather than two crossings as in [21] (heavier tails only).

We are interested here in a simple measure that simultaneously responds to the properties of peakedness and heaviness of tail of a density, a sort of measure of curvature, which is different from a measure of variance, dispersion, scale, or asymptotic order. In the following we derive such a measure within a framework called relative convexity. A similar idea was given by Beckenbach [7], and according to [23] also by E. Hopf (1926). H. Oja [20] uses similar concepts to define moment based kurtosis criteria. The measure we derive here is in fact instantaneous, but a basic inequality with respect to the Gaussian density is satisfied uniformly by almost all common symmetric densities. The form of the measure is similar to that of the geometric measure of curvature, and to the instantaneous kurtosis measure developed in [24]. In this paper we shall consider only symmetric, unimodal densities.

2.2. Relative convexity

Let $f:\mathbb{R}\to\mathbb{R}$ be increasing on the interval (a,b). The basic criterion for convexity of f is $f(\alpha x+\bar{\alpha}y)\le \alpha x+\bar{\alpha}y$ for all $x,y\in(a,b)$ [25, 26]. This may be interpreted as asserting that for any two points x and y in (a,b), the function value at all intervening points is less than the value of the linear function defined to match the value of f at the points x and y. In the intervals (a,x) and (y,b), the value of the convex function f will be greater than that of the linear function [26]. Analytically, we have f convex on (a,b) if,

$$f(y)\le f(x_0)+\frac{f(x_1)-f(x_0)}{x_1-x_0}(y-x_0)\quad \forall\, x_0\le y\le x_1$$

The inequality is reversed for y in (a,x_0), or (x_1,b). Letting $x_1\to x_0\equiv x$, we have the ordinary definition of convexity for differentiable f,

$$f(y)\ge f(x)+f'(x)(y-x)\quad \forall\, x,y\in(a,b)$$

Concavity can be defined similarly by reversing the inequalities. Thus convexity of a function on an interval can be seen as a relationship between the function and a linear model of the function based on the function value and first derivative.

This conception of convexity as a relationship between functions can be generalized to compare a function to nonlinear functions as well. This idea was also proposed by E. F. Beckenbach [7], and is related to the generalized convexity framework given in [23]. Let $h:\mathbb{R}\to\mathbb{R}$ be strictly increasing on (a,b). From considerations similar to those given in the linear case, we can define f to be *convex relative to* h on (a,b) if a model of h using an affine transform of f, given by $\alpha f+\beta$, defined so that f and h are equal at two given points, behaves in a manner similar to the line in the convex case. For any three points $x_0<y<x_1$ in the interval (a,b), we then have,

$$f(y)\le f(x_0)+\frac{f(x_1)-f(x_0)}{h(x_1)-h(x_0)}\left(h(y)-h(x_0)\right)\quad \forall y\in(x_0,x_1)$$

with the inequality reversed in (a,b) outside (x_0,x_1). Again letting $x_1\to x_0\equiv x$, we have f convex relative to h on (a,b) if,

$$f(y)\ge f(x)+\frac{f'(x)}{h'(x)}\left(h(y)-h(x)\right)\quad \forall\, x,y\in(a,b)\qquad (2.1)$$

We define relative concavity in the same way, reversing the inequalities. It may be verified that the condition (2.1) is equivalent to the differential definition of the convexity of the composite function $f\circ h^{-1}$ on the interval $(h(a),h(b))$. Thus we have f convex relative to h on (a,b) if $f\circ h^{-1}$ is convex on $(h(a),h(b))$, which is equivalent to $h\circ f^{-1}$ concave on $(f(a),f(b))$. The convexity of $f\circ h^{-1}$ is also equivalent in the sense of [23] to order 2 convexity of f with respect to h [23, p. 416]. According to [20], a similar concept is given by W. R. van Zwet (1964).

It is not difficult to show that the relative convexity relation induces a partial ordering on the set of functions increasing on (a,b) even without the requirement of differentiability [27]. Thus we can write $f\succeq h$ for f convex relative to h (h concave relative to f), and $f\preceq h$ for f concave relative to h (h convex relative to f). For f and h decreasing, we define $f(x)$ to be convex relative to $h(x)$ on (a,b) if $f(-x)$ is concave relative to $h(-x)$ on $(-b,-a)$.

If f and h are twice differentiable on (a,b), we can use the second derivative criterion for convexity to derive a simple criterion for $f\succeq h$. It may be verified that the condition,

$$D^2[f\circ h^{-1}](x)\ge 0\quad \forall\, x\in(h(a),h(b))$$

for f and h increasing, is equivalent to,

$$\frac{f''(x)}{f'(x)}\ge\frac{h''(x)}{h'(x)}\quad \forall\, x\in(a,b)\qquad (2.2)$$

This suggests that we can define a measure of relative convexity of one-dimensional increasing functions by the operator $K:f\to f''/f'=D\log Df$, such that $f\succeq h$ corresponds to $K(f)\ge K(h)$. The magnitude of this measure may be seen as a measure of function curvature, or of function order. Note that the relation is invariant to affine scaling of f, i.e. $K(f)=K(\alpha f+\beta)$ for all $\alpha>0$ and $\beta\in\mathbb{R}$.

2.3. Square-concavity and super-gaussianity

We define f to be *square-convex* if $f\succeq x^2$, and *square-concave* if $f\preceq x^2$ in the partial ordering defined by the relative convexity relation. In the important case of square-concavity, for f symmetric and strictly increasing on $(0,\infty)$, (2.1) becomes,

$$f(y)\le f(x)+\frac{f'(x)}{2x}(y^2-x^2)\ \ \forall\, x\ne 0, y\in\mathbb{R}\qquad (2.3)$$

and (2.2) becomes,

$$\frac{f''(x)}{f'(x)}\le\frac{1}{x}\quad \forall\, x>0\qquad (2.4)$$

We can apply the notion of square-concavity to densities that are symmetric, zero mean, and unimodal by taking the criterion for super-gaussianity to be negative log square-concavity on $(0,\infty)$. Then $p(x)$ will be super-gaussian if

$$D\log D\log\frac{1}{p(x)}\le\frac{1}{x}\quad \forall\, x>0\qquad (2.5)$$

In this form as an operator on densities, the measure can be seen as a sort of second order score function. Equivalently, $p(x)$ is negative log square-concave if it satisfies $p''(x)/p'(x)-p'(x)/p(x)\le 1/x$ for all $x>0$.

In [20] and references therein, two distribution functions F and G are compared by considering the convexity of the function $G^{-1}F$. This is obviously very similar to the criterion given here, but taking the second derivative of this expression and deriving a measure from an inequality similar to (2.2), we would get a measure in terms of the inverse of the distribution function. The inverse distribution can be used if it is available, but for many common densities it cannot be evaluated in closed form, notably for the Gaussian density. In [20], moment-based criteria are primarily considered for the measure of kurtosis. The advantage of the measure we derive is that it depends only on derivative information about the densities themselves, and contains no inverse distributions or moments.

A criterion essentially the same as the negative log square-concavity criterion is employed in the Bayesian image reconstruction literature [2, 3] to identify "edge-preserving" negative log priors. There the admissible negative log prior terms ϕ are such that $\phi(\sqrt{\cdot})$ is concave, which is equivalent to the square-concavity of ϕ as given here. Other similar criteria for half-quadratic algorithms are used to ensure the ϕ is at most square order in some sense [3]. The half-quadratic algorithm in [2] is in fact very similar to the algorithm for source estimation derived in the next section. The square-concavity idea is also used in [19, p. 52] and followed by [13, p. 2530]. We will use the square-concavity inequality (2.3) to prove descent for the component analysis algorithms given in section 4.

3. MAP ESTIMATION OF SUPER-GAUSSIAN SOURCES

Consider the MAP estimate of the sources in the linear model (1.1) for known A,

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

where $\bar{a}_j$ is the jth row of A, $f_i(x_i) \equiv -\log p_{X_i}(x_i)$ and $d_j(y_j - \bar{a}_j^T x) \equiv -\log p_{Y_j|X}(y_j|x)$. We assume that all source and noise random variables are independent with (not necessarily identical) unimodal, zero mean, super-gaussian densities. Defining $e \equiv y - Ax$, the problem can be written,

$$\hat{x} = \arg\min_{x,e} \sum_{i=1}^n f_i(x_i) + \sum_{j=1}^m d_j(e_j) \;\; s.t. \;\; Ax + e = y \quad (3.1)$$

Then defining $\tilde{A} \equiv [A \;\; I]$ and $\tilde{x} \equiv \left[x^T \; e^T\right]^T$, we have

$$\hat{\tilde{x}} = \arg\min_{\tilde{x}} \sum_{i=1}^{n+m} \tilde{f}_i(\tilde{x}_i) \quad s.t. \quad \tilde{A}\tilde{x} = y \quad (3.2)$$

where we define $\tilde{f}_i$ to range over the f_i and d_j functions, each of which is assumed square concave (this includes Gaussian noise and L_1 error as special cases). The formulation of the problem (3.2) includes the zero noise limit case for complete and overcomplete A, as well as the case for undercomplete A when noise dominates or source priors are uninformative. $\tilde{A}$ is overcomplete and full rank regardless of the dimension and rank of A. Thus we can solve all of the estimation problems mentioned by deriving an algorithm to solve the overcomplete case, $\hat{x} = \arg\min_x \sum_{i=1}^n f_i(x_i)$ such that $Ax = y$.

By assumption, we have each component function $f_i : \mathbb{R} \to \mathbb{R}$ symmetric, square-concave, and increasing with the magnitude of its argument. This 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 defined by $Ax = b$. We can use the inequality (2.3) to define a descent algorithm as follows. At each iteration l, for each $x_k, k = 1 \ldots N$, we have for arbitrary z,

$$\begin{aligned}f(z) - f(x_k) &= \sum_{i=1}^n f_i(z_i) - f_i(x_{i,k}) \\ &\leq \frac{1}{2}\nabla f(x_k)^T W_k^{-1}\left(z^2 - x_k^2\right) \\ &\equiv \frac{1}{2}\, z^T \Pi_k\, z - \frac{1}{2}\, x_k \Pi_k\, x_k \qquad (3.3)\end{aligned}$$

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

$$x_{k+1} \leftarrow \arg\min_{x\in C} x^T \Pi_k\, x \quad (3.4)$$

we can guarantee that right side of (3.3) is negative, and thus $f(x_{k+1}) \leq f(x_k)$. When C is the linear variety defined by $Ax = y$, the minimization can be carried out by solving,

$$\begin{bmatrix}\Pi_k & A^T \\ A & 0\end{bmatrix}\begin{bmatrix}x_{k+1} \\ \lambda\end{bmatrix} = \begin{bmatrix}0 \\ y\end{bmatrix} \quad (3.5)$$

where $\Pi_k(x_k) \equiv \mathrm{diag}(W_k^+ \nabla f(x_k))$, and W_k^+ is the pseudoinverse, obtained in this case by inverting the non-zero components of $\mathrm{diag}(x_k)$. We can write x_{k+1} in closed form as,

$$x_{k+1} = \Pi_k^+(x_k) A^T \left(A\,\Pi_k^+(x_k) A^T\right)^{-1} y \quad (3.6)$$

In the overcomplete case with noise, optimizing $[x^T e^T]^T$, we have,

$$x_{k+1} = \Pi_k^+(x_k) A^T \left(A\,\Pi_k^+(x_k) A^T + \Pi_k^+(e_k)\right)^{-1} y \quad (3.7)$$

where $V_k \equiv \mathrm{diag}(e_k)$ and $\Pi_k(e_k) \equiv \mathrm{diag}(V_k^+ \nabla d(e_k))$. It is unnecessary to solve for e_{k+1} as it is constrained to be $y - Ax_{k+1}$. In the undercomplete case, using the matrix inversion lemma, (3.7) becomes,

$$x_{k+1} = \left(\Pi_k(x_k) + A^T \Pi_k(e_k) A\right)^{-1} A^T \Pi_k(e_k) y \quad (3.8)$$

For super-gaussian priors, some components of the Π_k matrix may tend to infinity. In (3.7) this is unproblematic as Π_k^+ is used, in which the corresponding component will tend to zero. In (3.8) it may be necessary to explicitly check for very small components.

The descent properties guaranteed by these iterations play a role in the derivation of descent for component estimation algorithms given in the next section.

4. MAP ESTIMATION OF BASIS VECTORS

We now address the problem of estimating both A and $\mathbf{X}$. We first derive a descent algorithm, showing that the iterates decrease the negative log likelihood at each iteration. We approach the unconstrained problem by formulating it as a constrained optimization problem, and deriving a descent algorithm for the objective function such that the constraints are satisfied at each iteration. The objective function is precisely the posterior likelihood, not an approximation.

Let A be uniformly distributed over a compact set $\mathcal{S}_A \subset \mathbb{R}^{m\times n}$. Let $\mathbf{Y} = [y_1 \dots y_N]$ be observations from the linear model, $y_k = Ax_k + \nu_k$, with all random variables in the model independent, zero mean, symmetric, and negative log square-concave (super-gaussian). We consider the zero noise limit problem (3.1), with the general problem handled almost identically. We wish to find A and $\mathbf{X} = [x_1 \dots x_N]$ to minimize $\sum_k f(x_k) \equiv f(\mathbf{X})$ subject to $A\mathbf{X} = \mathbf{Y}$, $A \in \mathcal{S}_A$.

We develop a descent algorithm for $f(\mathbf{X})$ such that at each iteration we have $A\mathbf{X} = \mathbf{Y}$, and $A \in \mathcal{S}_A$. The iterates of the algorithm shall be denoted by $l = 1, 2, \dots$, and we denote $\hat{A}$, $\hat{\mathbf{X}}$, and $\hat{x}_k$ at the lth iteration by A_l, $\mathbf{X}_l$, and $x_{k,l}$ respectively. Considering the development of the preceding section, we shall assume that given A_{l+1}, we update $\hat{\mathbf{X}}$ according to (3.6), so that,

$$x_{k,l+1} = \Pi_{k,l}^{+} A_{l+1}^T (A_{l+1} \Pi_{k,l}^{+} A_{l+1}^T)^{-1} y_k \tag{4.1}$$

where $\Pi_{k,l} \equiv W_{k,l}^{+} \nabla f(x_{k,l})$, and $W_{k,l} \equiv \mathrm{diag}(x_{k,l})$. This update guarantees feasibility of $\mathbf{X}_{l+1}$ at the end of an iteration consisting of updating $\hat{A}$ and then updating $\hat{\mathbf{X}}$, and guarantees that $\hat{\mathbf{X}}$ moves in a reasonable direction. Although updating the source estimates according to (3.6) provides descent in the case of known constant A, (4.1) does not in general reduce $f(x_k)$, since (3.6) only guarantees that $f(x_{k,l+1}) < f(x_{k,l})$ for $x_{k,l}$ feasible. After A_l is updated to A_{l+1}, $\mathbf{X}_l$ is no longer feasible. Thus updating it according to (4.1) does not alone guarantee descent of $f(\mathbf{X}; A)$. We can, however, guarantee that $f(\mathbf{X}_{l+1}; A_{l+1}) < f(\mathbf{X}_l; A_l)$ by exploiting our knowledge that X_l will change to X_{l+1} according to (4.1) after we update A_l to A_{l+1}. Using the inequality (3.3) for square-concave functions, we have for each x_k,

$$f(x_{k,l+1}) - f(x_{k,l}) \le \frac{1}{2}\, x_{k,l+1}^T \Pi_{k,l}\, x_{k,l+1} - \frac{1}{2}\, x_{k,l} \Pi_{k,l} x_{k,l}$$

Using (4.1), for each x_k we have,

$$x_{k,l+1}^T \Pi_{k,l} x_{k,l+1} = y_k^T (A_{l+1} \Pi_{k,l}^{+} A_{l+1}^T)^{-1} y_k$$

Define the functions,

$$h_k(A) \equiv y_k^T (A \Pi_{k,l}^{+} A^T)^{-1} y_k - x_{k,l} \Pi_{k,l} x_{k,l}$$

and $h(A) = \sum_k h_k(A)$. Then we have $f(\mathbf{X}_{l+1}) - f(\mathbf{X}_l) \le h(A_{l+1})$. Now consider the value of h at A_l. We have, $h_k(A_l) = y_k^T (A_l \Pi_{k,l}^{+} A_l^T)^{-1} y_k - x_{k,l} \Pi_{k,l} x_{k,l} = \bar{x}_{k,l} \Pi_{k,l} \bar{x}_{k,l} - x_{k,l} \Pi_{k,l} x_{k,l}$, where $\bar{x}_k = \arg\min_x x^T \Pi_{k,l} x \ s.t.\ A_l x = y_k$. Thus,

$$h(A_l) = \sum_k h_k(A_l) = \sum_k \Big(\bar{x}_{k,l} \Pi_{k,l} \bar{x}_{k,l} - x_{k,l}^T \Pi_{k,l} x_{k,l} \Big) \le 0$$

since $A_l x_{k,l} = y_k$ for all k, and $\bar{x}_k$ achieves the minimum of $x^T \Pi_{k,l} x$ for $A_l x = y$. Thus if we can find $A_{l+1} \in \mathcal{S}_A$ such that $h(A_{l+1}) < h(A_l)$, then we will have,

$$f(\mathbf{X}_{l+1}) - f(\mathbf{X}_l) \le h(A_{l+1}) < h(A_l) \le 0$$

For simplicity, we use the method of gradient descent to decrease h. First suppose $\mathcal{S}_A$ is the sphere of unit Frobenius norm matrices. We could define geodesic gradient descent algorithms to maintain A in $\mathcal{S}_A$, but a more convenient method is to simply project the gradient of A evaluated at A_l onto the the subspace orthogonal to the "vector" A_l, i.e. onto the hyperplane tangent to the sphere of constant Frobenius norm at A_l [12]. Since the projection operator is positive semidefinite, the projected gradient is still a descent direction. For the projected gradient, we have,

$$\begin{aligned} A_{l+1} &= A_l - \alpha \,\mathrm{Proj}\,\Big(\frac{\partial h(A_l)}{\partial A}\Big) \\ &= A_l - \alpha \left(\frac{\partial h(A_l)}{\partial A} - \frac{\left\langle \frac{\partial h(A_l)}{\partial A}, A_l \right\rangle}{\langle A_l, A_l \rangle} A_l \right) \end{aligned} \tag{4.2}$$

The partial derivative of $h_k(A) = y_k^T (A \Pi_{k,l}^{+} A^T)^{-1} y_k$ with respect to A can be found using standard matrix calculus to be,

$$\frac{\partial h(A)}{\partial A} = -2 \sum_k (A \Pi_{k,l}^{+} A^T)^{-1} y_k y_k^T (A \Pi_{k,l}^{+} A^T)^{-1} A \Pi_{k,l}^{+} \tag{4.3}$$

With this we have for the inner product in (4.2),

$$\begin{aligned} \left\langle \frac{\partial h(A_l)}{\partial A}, A_l \right\rangle &= \mathrm{tr}\left(\frac{\partial h(A_l)}{\partial A} A_l^T \right) \\ &= -2 \sum_k \mathrm{tr}\Big((A_l \Pi_{k,l}^{+} A_l^T)^{-1} y_k y_k^T \Big) \\ &= -2 \sum_k y_k^T (A_l \Pi_{k,l}^{+} A_l^T)^{-1} y_k \end{aligned}$$

Define $\bar{\lambda}_{k,l} \equiv (A_l \Pi_{k,l}^{+} A_l^T)^{-1} y_k$. Then, absorbing the factor of 2 into the parameter α, and using the definition of $\bar{x}_k \equiv \Pi_{k,l}^{+} A_l^T \bar{\lambda}_{k,l}$, the update (4.2) becomes,

$$A_{l+1} = \left(1 - \alpha \frac{\sum_k \bar{\lambda}_{k,l}^T y_k}{\|A\|_F^2} \right) A_l + \alpha \sum_k \bar{\lambda}_{k,l} \bar{x}_{k,l}^T$$

Or, redefining α, we can write,

$$A_{l+1} = (1-\alpha) A_l + \alpha \, \|A\|_F^2 \, \frac{\sum_k \bar{\lambda}_{k,l} \bar{x}_{k,l}^T}{\sum_k \bar{\lambda}_{k,l}^T y_k} \tag{4.4}$$

which makes the setting of α much easier by reducing or eliminating dependence of the step size on the problem size. Note that in this setup we must finish the iteration by updating $\mathbf{X}_l$ according to (4.1), which amounts to a sort of double iteration on $\hat{\mathbf{X}}$. The iterations, however, must be done as prescribed for the algorithm to perform as stated. For example, we cannot simply iterate (4.1) twice since the derived descent depends on the two updates being done with different Π parameters (see (4.5) below).

It may be seen that in the case of complete A, this algorithm is equivalent to the Infomax algorithm of [16], and is thus a generalization of Infomax different from the algorithm given in [6]. The latter algorithm and the algorithm given in [12] may be seen as using alternative estimates of the Lagrange multiplier type vectors λ_k. Writing the Lagrangian for the optimization problem under consideration, we have,

$$L(A, \mathbf{X}) = \sum_k \Big(f(x_k) + \lambda_k^T (y_k - Ax_k) \Big) + \mu(\|A\|_F^2 - 1)$$

Setting the partial gradients of L equal to zero suggests,

$$\lambda_k^* = (A^* \Pi^+(x_k^*) A^{*T})^{-1} y_k$$

$$A^* = \frac{\sum_k \lambda_k^* x_k^{*T}}{\sum_k \lambda_k^{*T} y_k} \qquad x_k^* = \Pi^+(x_k^*) A^{*T} \lambda_k^*$$

Thus, except for the norm multiplier, (4.4) can be seen as a fixed point Lagrangian algorithm. Indeed, eliminating $\|A\|_F^2$ from the expression yields an algorithm that is shown experimentally to converge to a solution with unit Frobenius norm, though it may temporarily increase the objective function on its way. The iteration given in (4.4), if initialized with a unit norm matrix, will, as expected, steadily but very slightly increase the norm of A, usually coming to a fixed point before increasing by more than 10^{-4}. Also as expected it monotonically decreases the objective function for "natural" step sizes of approximately 10^{-2} or 10^{-1}.

One problem with algorithms of this form for component estimation however, as noted in [6] and elsewhere, is that there is nothing to prevent individual columns from going to zero, reaching a sort of degenerate solution. We can eliminate this problem by constraining A to have unit column norm. In this case the Lagrangian is,

$$L(A, \mathbf{X}) = \sum_k \Big(f(x_k) + \lambda_k^T (y_k - A x_k) \Big) + \sum_{i=1}^n \mu_i (a_i^T a_i - 1)$$

where $A = [a_1 \dots a_n]$. Following the example of the constrained Frobenius norm case, we can derive a fixed point Lagrangian algorithm that converges to an A^* with unit column norms. We follow the same protocol suggested by the globally convergent algorithm. Let $W_{k,l} = \text{diag}(x_{k,l})$. We first calculate the Lagrange multiplier and $\bar{x}_k$ estimates,

$$\Pi_{k,l} = \text{diag}(W_{k,l}^+ \nabla f(x_{k,l})) \qquad \bar{\lambda}_{k,l} = (A_l \Pi_{k,l}^+ A_l^T)^{-1} y_k$$

$$\bar{\mu}_l = \sum_k W_{k,l} A_l^T \bar{\lambda}_{k,l} \qquad \bar{x}_{k,l} = \Pi_{k,l}^+ A_l^T \bar{\lambda}_{k,l}$$

Then update A,

$$A_{l+1} = (1 - \alpha)\, A_l + \alpha \bigg(\sum_k \bar{\lambda}_{k,l} \bar{x}_{k,l}^T \bigg) \text{diag}(\bar{\mu}_l)^{-1}$$

Finally, update the x_k using the new A, and (old) $\Pi_{k,l}$.

$$\lambda_{k,l} = (A_{l+1} \Pi_{k,l}^+ A_{l+1}^T)^{-1} y_k \qquad x_{k,l+1} = \Pi_{k,l}^+ A_{l+1}^T \lambda_{k,l} \quad (4.5)$$

To conclude, we give a duality relationship between sub- and super-gaussian densities as here defined, which allows similar algorithms to be formulated for sub-gaussian estimation. The Fenchel-Legendre conjugate of a convex function f, defined by $f^*(\phi) = \sup_x \phi x - f(x)$, is used in the dual of a convex optimization problem, e.g. $\min_{Ax=y} f(x)$ can be obtained by solving $\max_{V\phi=0} \bar{x}^T \phi - f^*(\phi)$, where V is a basis for the null space of A, and $A\bar{x} = y$ (see[25]). For f concave, $f^*(\phi) = \inf_x \phi x - f(x)$.

Theorem 1 (Fenchel-Legendre conjugate symmetry [27]). *Let f be increasing and convex or concave on $\mathcal{D}$, with Fenchel conjugate f^* defined on $\mathcal{D}^*$. Then $f \succeq x^2$ on $\mathcal{D}$ if and only if $f^* \preceq \phi^2$ on $\mathcal{D}^*$. Also, $f \preceq \log$ on $\mathcal{D}$ if and only if $f^* \succeq \log$ on $\mathcal{D}^*$.*

This is similar to the result that an increasing function f is convex on (a, b) if and only if f^{-1} is concave on $(f(a), f(b))$.

5. REFERENCES

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal of Scientific Computation*, vol. 20, no. 1, pp. 33–61, 1998.

[2] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 3, pp. 367–383, 1992.

[3] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization and FFT's," *IEEE Trans. Image Processing*, vol. 11, pp. 803–851, 1999.

[4] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, pp. 803–851, 1999.

[5] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," *Network: Computation in Neural Systems*, vol. 7, pp. 333–339, 1996.

[6] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, pp. 337–365, 2000.

[7] E. F. Beckenbach, "Generalized convex functions," *Bull. Am. Math. Soc.*, vol. 43, pp. 363–371, 1937.

[8] D. J. Field, "What is the goal of sensory coding?," *Neural Computation*, vol. 6, pp. 559–601, 1994.

[9] M. S. Lewicki and B. A. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *J. Opt. Soc. Am. A*, vol. 16, no. 7, pp. 1587–1601, 1999.

[10] A. Hyvärinen, "Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood," *Neurocomputing*, vol. 22, pp. 49–67, 1998.

[11] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley and Sons, Ltd., West Sussex, England, 2002.

[12] K. Kreutz-Delgado and B. D. Rao, "Focuss-based dictionary learning algorithms," in *Wavelet Applications in Signal and Image Processing VII: Proc. of SPIE*. SPIE, 2000, vol. 4119.

[13] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Computation*, vol. 13, pp. 2517–2532, 2001.

[14] H. Lu, Y. Fainman, and R. Hecht-Nielsen, "Image manifolds," in *Applications of Artificial Neural Networks in Image Processing III: Proc. of SPIE*. SPIE, 1998, vol. 3307.

[15] P. Comon, "Independent component analysis: A new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

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

[17] T. Lee, *Independent Component Analysis*, Kluwer Academic Publishers, 1998.

[18] M. Girolami, *Self-Organizing Neural Networks: Independent Component Analysis and Blind Source Separation*, Springer, 1999.

[19] T. S. Jaakola, *Variational Methods for Inference and Estimation in Graphical Models*, Ph.D. thesis, Massachusetts Institute of Technology, 1997.

[20] H. Oja, "On location, scale, skewness and kurtosis," *Scandinavian Journal of Statistics*, vol. 8, pp. 154–168, 1981.

[21] A. Mansour and C. Jutten, "What should we say about the kurtosis?," *IEEE Signal Processing Letters*, vol. 6, no. 12, pp. 321–322, 1999.

[22] H. M. Finucan, "A note on kurtosis," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, no. 1, pp. 111–112, 1964.

[23] S. Karlin and W. J. Studden, *Tchebycheff Systems: With Applications in Analysis and Statistics*, Interscience, New York, 1966.

[24] P. J. Loughlin and K. L. Davidson, "Instantaneous kurtosis," *IEEE Signal Processing Letters*, vol. 7, no. 6, pp. 156–159, 2000.

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

[26] J. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms I*, Sringer-Verlag, 1993.

[27] J. A. Palmer, "Function curvature, relative convexity, and conjugate curvature," Tech. Rep., ECE Dept., UCSD, 2002.