Convex/Schur-Convex (CSC) Log-Priors and Sparse Coding

K. Kreutz-Delgado, B.D. Rao, and K. Engan*
Electrical and Computer Engineering Department
Jacobs School of Engineering
University of California, San Diego
{kreutz, brao, kengan}@ucsd.edu

T.-W. Lee and T. Sejnowski
Computational Neurobiology Laboratory
The Salk Insitute, La Jolla, California
{tewon, terry}@salk.edu

Abstract

The "folk theorem" that sparsity inducing priors should be supergaussian can be rigorously stated in the low-noise limit, assuming the validity of a particular stochastic generative model. For the assumed model it is shown that supergaussianess is necessary, but not sufficient, for sparse signal coding when a *Maximum A Posterior* (MAP) coding is found.

1 Introduction

It has been noted by a variety of investigators in several different research domains (human vision, signal processing, etc) that a generative model appropriate for understanding sparse coding and Independent Component Analysis (ICA) is given by the system of equations,

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

where y is an observed signal vector, x is an unobserved ("blind") source vector, the columns of A form an overcomplete dictionary, and ν is a measurement noise vector which is independent of the source.

The system (1) is profitably interpreted within a Bayesian framework. The components, $x[k]$, $k = 1, \cdots n$, of the source vector, x, are independent components within this framework when the statistical prior, $P(x)$, factors into a product of marginal probabilities as $P(x) = P(x[1]) \cdots P(x[n])$. Within this ICA model an estimate of the source vector x is said to provide a "factorial code" of the measured signal y. This shows the relationship between

*Visiting Scholar. Current address: Dept. of ECE, Høgskolen i Stavanger, Stavanger Norway. Email: Kjersti@hsr.no

ICA and factorial coding. It is generally believed that for x to provide a sparse coding of y, the marginal prior $P(x[k])$ should be supergaussian. This fact has not been proven in a strictly rigorous manner, but has a plausible intuitiveness and has been demonstrated in a variety of simulations and applications. It serves as a useful "folk theorem" for obtaining sparse codes via the model (1).

We make rigorous statements concerning when a *maximum a posteriori* (MAP) estimate of the source x will provide a sparse coding of the measured signal y in the low-noise limit. We show that a requirement for obtaining sparse codings is that the negative log–prior, $-\log(P(x))$, be concave. We also argue that an additional desirable property is that $-\log(P(x))$ be Schur-concave. A rigorous theoretical justification of the results discussed in this paper can be found in the report [4].

2 Bayesian Framework

Stochastic Generative Model. A Bayesian interpretation is obtained from the generative signal model (1) by assuming that x has the parameterized (generally non–gaussian) pdf,

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

with parameter vector p. Similarly, the noise ν is assumed to have the parameterized (possibly nonguassian) density $P_q(\nu)$ with parameter vector q. It is assume that x and ν have zero means and that their densities obey the property $d(x) = d(|x|)$, for $|\cdot|$ defined component-wise. This is equivalent to assuming that the densities are symmetric with respect to sign changes in the components of x, $x[i] \leftarrow -x[i]$, and therefore that the skews of these densities are zero. Here, with a slight abuse of notation, we allow the differing subscripts q and p to indicate that d_q and d_p may be *functionally* different as well as parametrically different. We will refer to densities like (2), for suitable additional constraints on $d_p(x)$ to be stated below, as *Hyper-Generalized Gaussian Distributions* (HGDs). As discussed below, they are a superset of the well-known generalized gaussian distributions (GGDs) (also known as exponential power distributions or Box-Tiao distributions) [3].

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

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

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

Given an observation, y, maximizing (3) with respect to x yields the MAP estimate $\widehat{x}$. This ideally results in a sparse coding of the observation, a requirement which places functional constraints on the probability density functions. Note that β is independent of x and can be ignored when optimizing (3).

The MAP estimate equivalently is obtained from minimizing the the negative logarithm of $P(x|y)$, which is (to within irrelevant multiplicative and additive constants) equal to the problem,

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

where $\lambda = \gamma_p / \gamma_q$, and $d_q(y - Ax) = d_q(Ax - y)$ by our assumption of symmetry. Note that $\lambda \to 0$ as $\gamma_p \to 0$ which (consistant with the generative model (1)) we refer to as the

low noise limit. Because the mapping A is assumed to be onto, in the low noise limit the optimization (5) is equivalent to the linearly constrained problem,

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

It is evident that the structure of $d_p(\cdot)$ is critical for obtaining a sparse coding, $\widehat{x}$, of the observation y.

Independent Component Analysis (ICA). An important class of densities is given by the *generalized gaussians* for which

$$d_p(x) = ||x||_p^p = \sum_{k=1}^{n} |x[k]|^p\,, \tag{7}$$

for $p > 0$ [3]. This is a special case of the ℓ_p class (the "p-class") of functions which allow p to be negative which is discussed in [9, 4]. Note that this function has the special property of *separability*,

$$d_p(x) = \sum_{k=1}^{n} d_p(x[k])\,,$$

which corresponds to *factorizability* of the density $P_p(x)$,

$$P_p(x) = \prod_{k=1}^{n} P_p(x[k])\,,$$

and hence to *independence of the components of* x. It is, in fact, now well-known that the assumption of independent components allows the problem of solving the generative model (1) for x to be interpreted as an Independent Component Analysis (ICA) problem [1, 8, 7, 10]. In [9] and [4] is developed a large class of parameterizable separable functions $g_p(x)$ consistent with the ICA assumption. Note that given such a class, it is natural to examine the issue of finding a best fit within this class to the "true" underlying prior density of x. This is a problem of parametric density estimation of the true prior where one attempts to find an optimal choice of the model density $P_p(x)$ by an optimization over the parameters p which define the choice of a prior from within the class. This is, in general, a difficult problem which may require the use of monte-carlo, evolutionary programming, and/or stochastic search techniques.

Sparsity Inducing Priors. Can the belief that supergaussian priors, $P_p(x)$, are appropriate for finding sparse solutions to (1) (see, e.g., [2, 7]) be clarified or made rigorous? It is well known that the generalized gaussian distribution arising from the use of (7) yields supergaussian distributions (positive kurtosis) for $p < 2$ and subgaussian (negative kurtosis) for $p > 2$. However, we will argue that the condition for obtaining sparse solutions in the low noise limit is the stronger requirment that $p \leq 1$, in which case the separable function $d_p(x)$ is *concave and Schur-concave*. This indicates that supergaussianess (positive kurtosis) alone is *necessary* but *not sufficient* for inducing sparse solutions. Rather, sufficiency is given by the requirement that $-\log P_p(x) \approx d_p(x)$ be Concave/Schur-Concave(CSC).

We've seen that the function $d_p(x)$ has an interpretation as a (negative logarithm of) a Bayesian prior *or* as a penalty function enforcing sparsity in (5) where $d_p(x)$ should serve as a "relaxed counting function" on the nonzero elements of x. Our perspective emphasizes the

fact that $d_p(x)$ serves *both* of these goals simultaneously. Thus, good regularizing functions, $d_p(x)$, should be flexibly parameterizable so that $P_p(x)$ can be optimized over the parameter vector p to provide a good parametric fit to the underlying environmental probability density function, *and* the functions should also have analytical properties consistent with the goal of enforcing sparse solutions. Such properties are discussed in the next section.

3 Majorization and Schur-Concavity [5]

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* [5, 12],

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

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 then defines the Lorentz order.

We are interested in scalar-valued functions of x which are consistent with majorization. Such functions 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) than $d(y)$ is *less than* $d(x)$ for $d(\cdot)$ Schur-concave.

- We assume that Schur-Concavity is 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 [11]

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

$\forall x, y \in R_+^n, \forall \gamma, 0 \leq \gamma \leq 1$. In addition, a scalar function is said to be permutation invariant 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*:

$$\text{Concavity} + \text{Permutation Invariance} \Rightarrow \text{Schur-Concavity}.$$

Now consider the low-noise sparse inverse problem (6). It is well known that subject to linear constraints, a concave function on R_+^n takes its minima on the *boundary* of R_+^n [11], and as a consequence these minima are therefore *sparse* (see Figure 1).

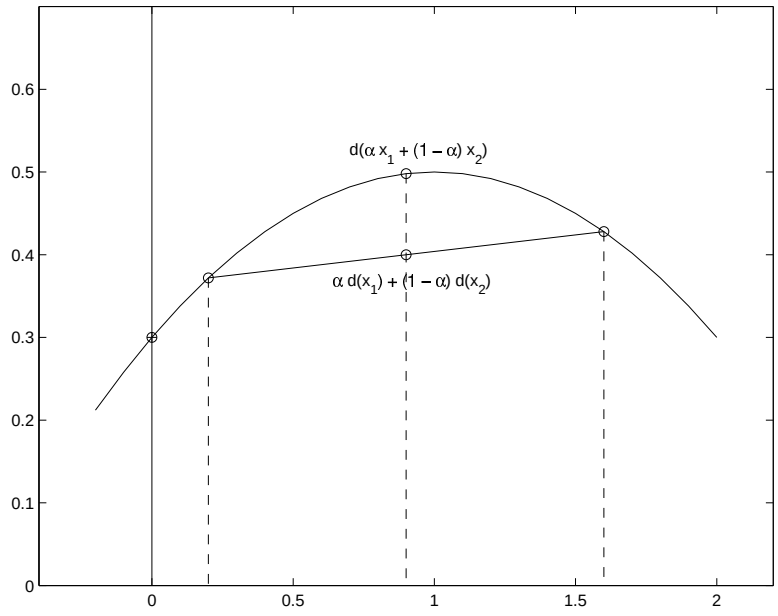

Figure 1: Minimization of Concave Function Yields Sparsity. Sliding down a bowl-shaped (concave) function results in a zero-component value on the boundary of an orthant.

- 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 [4] 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*. Recall that separability of $d_p(x)$ corresponds to *factorizability* of $P_p(x)$. Thus *separability* of $d(x)$ corresponds to the assumption of *independent components* of x under the model 1). 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) [6, 8, 10]. We have developed effective algorithms for solving the optimization problem (6) for sparse solutions when $d_p(x)$ is separable and concave [4, 9].

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). Unless care is taken, this relaxation can significantly complicate the analysis and development of optimization algorithms. Fortunately, it can be shown that the algorithms of [4, 9] can be applied to solve the low-noise DCA problem provided appropriates choice of non-separable diversity functions are made.

4 Supergaussian Priors and Sparse Coding

The P-class of diversity measures for $0 < p \leq 1$ result in sparse solutions to the low-noise coding problem (6). These separable and concave (and thus Schur-concave) diversity measures correspond to supergaussian generalized gaussian priors, consistent with the "folk theorem" that supergaussian priors are sparsity enforcing priors. However, taking $1 \leq p < 2$ results in supergaussian priors which are *not* sparsity enforcing. Taking p to be between 1 and 2 yields a $d_p(x)$ which is *convex*, and therefore *not* concave. This is consistent with the well-known fact that for this range of p, the p^{th}-root of $d_p(x)$ is a norm. Minimizing $d_p(x)$ in this case drives x towards the origin, favoring "concentrated" rather than "sparse" solutions. In Figure (2) is shown the level curves of $d_p(x)$ for several values of p along with

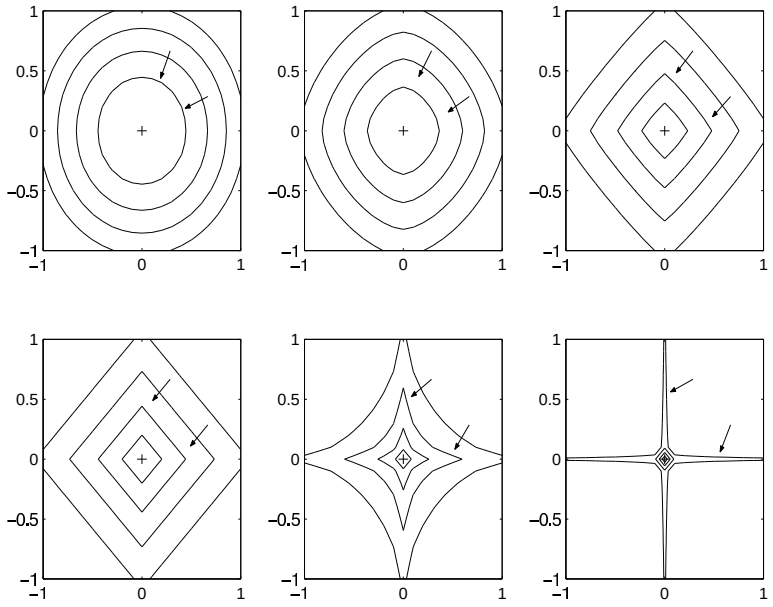

Figure 2: Contour Plots and Steepest Descent Vectors of $d_p(x)$ for Various Values of p. Note that for $p > 1$, the directions of steepest descent is directed towards the origin, while for $p \leq 1$ they are directed towards the boundary of the positive orthant enforcing sparse solutions.

the directions of steepest descent (negative gradient) of $d_p(x)$ at two fixed points in the positive orthant. It is evident that the gradients for $1 < p < 2$ point towards the origin, while the gradients for $0 < p \leq 1$ point towards the boundary of the orthant. We see that if a sparse coding is to be found based on obtaining a MAP estimate to the low-noise generative model (1) then, in a sense, supergaussianess is a necessary but not sufficient condition for a prior to be sparsity enforcing. A sufficent condition for obtaining a sparse MAP coding is that the negative log-prior be Concave/Schur-concave (CSC).

5 Conclusions

The majorization framework gives insight into diversity measures which are useful for enforcing sparse solutions to linear inverse problems. This has sharpened our understanding of the structure of sparsity-inducing priors, at least for the low-noise limit of the generative model (1). In particular we have seen that Concave/Schur-concave (CSC) negative log-priors function as good diversity and sparsity enforcing measures, and that separability of such measures is consistent with ICA generative models. We have seen that such priors include supergaussian priors, but that supergaussianess is not sufficient to ensure sparse codings via MAP estimation. It still remains to determine if this low-noise limitation holds for other sparsity-enforcing algorithms which have been proposed in the literature. To the degree that such algorithms can be shown to be variants of the MAP coding procedure, then this limitation should hold in the low-noise limit. In the non-low-noise case, thresholding may be required to separate signals from spurious noise components. In this case it may be undesirable to overly enforce sparse solutions prior to the thresholding stage.

References

[1] P. Comon, "Independent Component Analysis: A New Concept?" *Signal Proc.*, Vol. 36, pp. 287-314, 1994.

[2] D. Field, "What is the Goal of Sensory Coding," *Neural Comp.,* 6, 559-99, 1994.

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

[4] 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.

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

[6] 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.

[7] B.A. Olshausen, "Learning Linear, Sparse, Factorial Codes", September 1996.

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

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

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

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

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