An Improved FOCUSS-Based Learning Algorithm for Solving Sparse Linear Inverse Problems

Joseph F. Murray and Kenneth Kreutz-Delgado
Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407
jfmurray@ucsd.edu, kreutz@ece.ucsd.edu

Abstract

We develop an improved algorithm for solving blind sparse linear inverse problems where both the dictionary (possibly overcomplete) and the sources are unknown. The algorithm is derived in the Bayesian framework by the maximum a posteriori *method, with the choice of prior distribution restricted to the class of concave/Schur-concave functions, which has been shown previously to be a sufficient condition for sparse solutions. This formulation leads to a constrained and regularized minimization problem which can be solved in part using the FOCUSS (Focal Underdetermined System Solver) algorithm for vector selection. We introduce three key improvements in the algorithm: an efficient way of adjusting the regularization parameter, column normalization that restricts the learned dictionary, and reinitialization to escape from local optima.*

Experiments were performed using synthetic data with matrix sizes up to 64x128, and the algorithm is shown to solve the blind identification problem, recovering both the dictionary and the sparse sources. The improved algorithm is shown to be much more accurate than the original FOCUSS-Dictionary Learning algorithm when using large matrices.

We also test our algorithm on natural images, and show that a learned overcomplete representation can encode the data more efficiently than a complete basis at the same level of accuracy.

1 Introduction

In many applications a small number of sources drawn from a very large set of possible sources are mixed together to generate observed data. Examples include magnetoencephalography (MEG), where spatially localized signals in the brain mix with each other before reaching the sensors, leading to the related problems of localizing the sources and removing undesired artifacts [1, 2]. In visual processing, a few objects can be thought of as generating a large number of image pixels at the camera or retina. These problems can be modeled as an inverse problem with sparse solutions,

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

where y is the observed data, $A \in \mathbb{R}^{m\times n}$ is the generating matrix or dictionary (the columns of A are the dictionary vectors, and A may be overcomplete, i.e. $n > m$), x is the sparse solution (most of the elements of x are identically zero), and ν is additive noise. In this paper we present an improved version of the algorithm derived in [3] for solving (1) when both x and A are assumed to be unknown random variables. The problem of determining A (respectively x) when x (respectively A) is unknown is known as the blind identification problem. The algorithm is designed to learn an environmentally adapted dictionary A which is capable of representing signals as sparsely as possible with a minimum reconstruction error.

Sparse solutions to (1) can also be useful in the feature extraction stage (or concept generation stage) of pattern recognition problems, where a set of informative features is culled from the available data, which can improve the generalization performance of pattern recognition methods like support vector machines and neural networks [4, 5]. Other applications include data compression, high-resolution spectral estimation, direction-of-arrival estimation, speech coding, and function approximation; see the references in [6] for details.

Our algorithm is tested on the difficult nonlinear inverse problem of simultaneously correctly identifying the true values of the unknowns A and x (the blind identification problem). Using synthetic data, it is shown that the algorithm can recover a high percentage of the columns of A and (after adjusting for possible column permutations) original sources x. The advantage of using overcomplete dictionaries for encoding images is also shown. At the same level of reconstruction error the overcomplete representation is more efficient, i.e. lower entropy (bits/pixel).

Other approaches to solving the overcomplete problem (1) have been presented, including those based on independent component analysis (ICA) [7, 8], independent factor analysis [9], and the original FOCUSS-based learning algorithm [10].

2 Learning algorithm

The FOCUSS-based dictionary learning algorithm and convergence proofs are fully described in [3], and here we only present a basic outline sufficient to describe our recent improvements. Given a set of training data $Y = (y_1, .., y_N)$, we would like to solve (1) for the jointly random A and $X = (x_1, .., x_N)$ using maximum *a posteriori* estimation,

$$(\widehat{A}_{\text{MAP}}, \widehat{X}_{\text{MAP}}) = \arg \max_{A \in \mathcal{A},\, X} P(A, X|Y), \qquad (2)$$

where A is constrained to the compact (closed and bounded) set $\mathcal{A}$. Assuming the prior distribution of the sources x is a generalized Gaussian of the form,

$$P_p(x) = Z_p^{-1} e^{-\gamma_p\, d_p(x)}\,, \quad Z_p = \int e^{-\gamma_p\, d_p(x)} dx\,, \quad (3)$$

where the parameter p determines the shape of distribution, and Z_p is a normalizing constant to ensure $P_p(x)$ is a density function. Here, the function $d_p(x)$ is related to the p-norm-like measure,

$$d_p(x) = \|x\|_p^p = \sum_{i=1}^{n} |x[i]|^p\,, \quad 0 \le p \le 1\,, \qquad (4)$$

Sparsity is defined as the number of elements of x that are zero. When $p = 0$, $d_p(x)$ is a count of the number of non-zero elements of x, or *diversity*, where diversity $= n-$ sparsity. Other parametric forms of the prior $d_p(x)$ that lead to sparse solutions are discussed in [11].

The additive noise ν has a distribution $P_q(\nu)$ of the same form as (3), and here we assume the noise vector is Gaussian (i.i.d. with zero mean) so that $q = 2$ and $d_q(\nu) = \|\nu\|_2^2$. Using the prior distributions (3) and assuming that the observations Y are independent, (2) can be written (see eq. (23)-(25) in [3]),

$$(\widehat{A}_{\text{MAP}}, \widehat{X}_{\text{MAP}}) = \arg \min_{A \in \mathcal{A},\, X} \left\langle d_q(y - Ax) + \lambda\, d_p(x) \right\rangle_N\,, \qquad (5)$$

where the $\langle\cdot\rangle_N$ denotes averaging over the N vectors in the training set. Using the form of $d_p(x)$ and $d_q(x)$ in (4) this becomes,

$$(\widehat{A}_{\text{MAP}}, \widehat{X}_{\text{MAP}}) = \arg \min_{A \in \mathcal{A},\, X} \left\langle \|y - Ax\|_2^2 + \lambda\, \|x\|_p^p \right\rangle_N\,. \qquad (6)$$

This is a constrained optimization problem that attempts to simultaneously minimize the reconstruction error $\|y - Ax\|_2^2$ and maximize the sparsity of the solutions by minimizing $\|x\|_p^p$. The regularization parameter λ controls the trade-off between the allowed error and desired sparsity, and its choice is critical for finding accurate solutions, but unfortunately there is no theoretical way of choosing an optimal value.

The algorithm contains two major parts, a sparse vector selection step and a dictionary learning step. The Focal Underdetermined System Solver (FOCUSS) was designed to solve for sparse solutions of linear inverse problems when A is known [12, 6], and performs the vector selection step of the algorithm. The FOCUSS algorithm is not a linear transform of the data, even in the case where A is complete, because it maximizes the sparsity of the learned sources x, with the tradeoff that perfect reconstruction is generally not achieved even in the noise-free case. The dictionary learning (A update) step is performed by gradient descent. The iterations are given by

$$\begin{aligned}
\Pi_{\widehat{x}_k}^{-1} &= \text{diag}(|\widehat{x}_k[i]|^{2-p}) \\
\widehat{x}_k &\leftarrow \Pi_{\widehat{x}_k}^{-1} \widehat{A}^T \left(\lambda_k I + \widehat{A} \Pi_{\widehat{x}_k}^{-1} \widehat{A}^T\right)^{-1} y_k \\
\Sigma_{y\hat{x}} &= \frac{1}{N} \sum_k y_k \widehat{x}_k^T\,, \\
\Sigma_{\hat{x}\hat{x}} &= \frac{1}{N} \sum_k \widehat{x}_k \widehat{x}_k^T \\
\delta \widehat{A} &= \widehat{A}\, \Sigma_{\hat{x}\hat{x}} - \Sigma_{y\hat{x}} \\
\widehat{A} &\leftarrow \widehat{A} - \gamma \left(\delta \widehat{A} - \text{trace}\,(\widehat{A}^T \delta \widehat{A}) \widehat{A}\right)\,. \qquad (7)
\end{aligned}$$

where λ_k is the regularization parameter and $\gamma (> 0)$ controls the learning rate. The second line of (7) is the FOCUSS algorithm for updating the x_k solutions.

We now discuss the three key improvements in the algorithm: an efficient way of adjusting the regularization parameter λ_k, column normalization that restricts the learned $\widehat{A}$, and reinitialization to escape from local optima. The regularization parameter λ_k may be set independently for each vector in the training set, and a number of methods have been suggested, including quality-of-fit (which requires a certain level of reconstruction accuracy), sparsity (requiring a certain number of non-zero elements), and the L-curve which attempts to find an optimal tradeoff [10]. The L-curve method works well, but it requires solving a one-dimensional optimization for each λ_k which becomes computationally expensive for large problems. Alternatively, we use a heuristic method that allows the tradeoff between error and sparsity to be tuned for each application, while letting each training vector y_k have its own regular-

ization parameter λ_k to improve the quality of the solution,

$$\lambda_k = \lambda_{\max}\left(1 - \frac{\|y_k - \widehat{A}\widehat{x}_k\|}{\|y_k\|}\right), \quad \lambda_k, \lambda_{\max} > 0. \tag{8}$$

For data vectors that are represented accurately, λ_k will be large, driving the algorithm to find more sparse solutions. If the signal-to-noise ratio (SNR) can be estimated, we can set $\lambda_{\max} = (\text{SNR})^{-1}$.

When $q = 2$ (the case of Gaussian measurement noise), the algorithm converges to a local minimum of (6) (see the appendix of [3]) if $\widehat{A}$ is restricted to a bounded subset of $\mathbb{R}^{m\times n}$. This is accomplished by restricting $\widehat{A}$ to the set,

$$\mathcal{A} = \{\, A \mid \|A\|_F = 1 \,\} \subset \mathbb{R}^{m\times n} \tag{9}$$

where $\|A\|_F = \sqrt{\text{trace}(A^T A)}$ is the Frobenius matrix norm. To ensure this, $\widehat{A}$ is normalized to $\|\widehat{A}\|_F = 1$ by normalizing each of the columns a_i,

$$a_i \leftarrow \frac{a_i}{\sqrt{n}\,\|a_i\|_2}. \tag{10}$$

This keeps the norm of all columns equal, which is important when using diversity measures with $p > 0$, because these measures penalize terms with large magnitudes (they are only approximating the $p = 0$ measure of sparsity). If a column had a small relative magnitude, the weights of its coefficients would be large and it would be penalized more than a column with a larger norm. This leads to certain columns being underused, which is especially troublesome in the overcomplete case. The convergence proof in [3] requires only that $\|\widehat{A}\|_F = 1$, so it is important to show that restricting the evolution of $\widehat{A}$ to the subset of column normalized matrices allows us to reach to true solution A (see Appendix). In the Bayesian framework, the column normalization can be viewed as a more restrictive prior than using only the Frobenius norm constraint (as was done in the original algorithm).

The optimization problem (6) is concave when $p \leq 1$, so there will be multiple local minima. The FOCUSS algorithm is only guaranteed to converge to one of these local minima, but in some cases it is possible to determine when that has happened by noticing if the sparsity is too low. Periodically (after a large number of iterations) the sparsity of the solutions $\widehat{x}_k$ is checked, and if found too low, $\widehat{x}_k$ is reinitialized randomly. The algorithm is also sensitive to initial conditions and prior information may be incorporated into the initialization to help convergence to the global solution.

The learning algorithm in (7) is a combined iteration, meaning that the FOCUSS step is only allowed to run for one iteration (not until full convergence) before the A update step. This means that during early iterations, the $\widehat{x}_k$ are in general not sparse. To facilitate learning A, the covariances $\Sigma_{y\hat{x}}$ and $\Sigma_{\hat{x}\hat{x}}$ are calculated with sparsified $\widehat{x}_k$ that have all but the $\tilde{r}$ largest elements set to zero. The value of $\tilde{r}$ is usually set to the largest desired number of non-zero elements, but this choice does not appear to be critical.

3 Synthetic data experiments

To test the algorithm's ability to recover the true A and x_k solutions, experiments were conducted using synthetically generated data. Elements of A were drawn from a normal distribution with $\mu = 0, \sigma^2 = 1$ ($\mathcal{N}(0,1)$), and the matrix was normalized as in (10) to $||A||_F = 1$. Sparse source vectors were created with diversity r, and the value of each non-zero element $x_k[i]$ was also drawn from $\mathcal{N}(0,1)$ and limited so that $x_k[i] > 0.1$. The input data y_k were generated using (1) with no noise added. Matrix sizes were 20x30 and 64x128, and r was set to fixed values (4 and 7) and randomly (5..10 and 10..15).

The columns of the initial dictionary $\widehat{A}_{init}$ were set to the first n data vectors y_k. The $\widehat{x}_k$ are initialized to the pseudoinverse solution, $\widehat{x}_k = \widehat{A}^T(\widehat{A}\widehat{A}^T)^{-1}y_k$. The parameters were set as follows: $p = 1.0, \gamma = 1.0, \lambda_{\max} = 2\times 10^{-3}$ (low noise, assumed SNR $\approx$ 27 dB). The algorithm was run for 500 iterations through the entire data set, and during each iteration $\widehat{A}$ was updated after updating 100 data vectors $\widehat{x}_k$. The training vectors were presented in a random order each iteration.

As a measure of performance, we find the number of columns of A that were matched during learning. Because A can only be learned to within column permutations and sign and scale changes, the columns are normalized so that $||\widehat{a}_i|| = ||a_j|| = 1$ and $\widehat{A}$ is rearranged columnwise so that $\widehat{a}_j$ is given the index of the closest match in A (in the minimum 2-norm sense) . A match is counted if

$$1 - |a_i^T \widehat{a}_i| < 0.01. \tag{11}$$

Similarly, the number of matching $\widehat{x}_k$ are counted (after rearranging the elements in accordance with the indices of the rearranged $\widehat{A}$)

$$1 - |x_i^T \widehat{x}_i| < 0.05. \tag{12}$$

If the data is generated by an A that is not column normalized, other measures of performance need to be used to compare x_k and $\widehat{x}_k$.

The performance is summarized in Table 1, where the original FOCUSS-based dictionary learning algorithm (FOCUSS-DL) [3] is compared with the improved column-normalized algorithm presented here (FOCUSS-CNDL). For the 20x30 matrix 1000 training vectors were used, and for the 64x128 matrix 10,000 were used. Results are averaged over four or more trials. For the 64x128

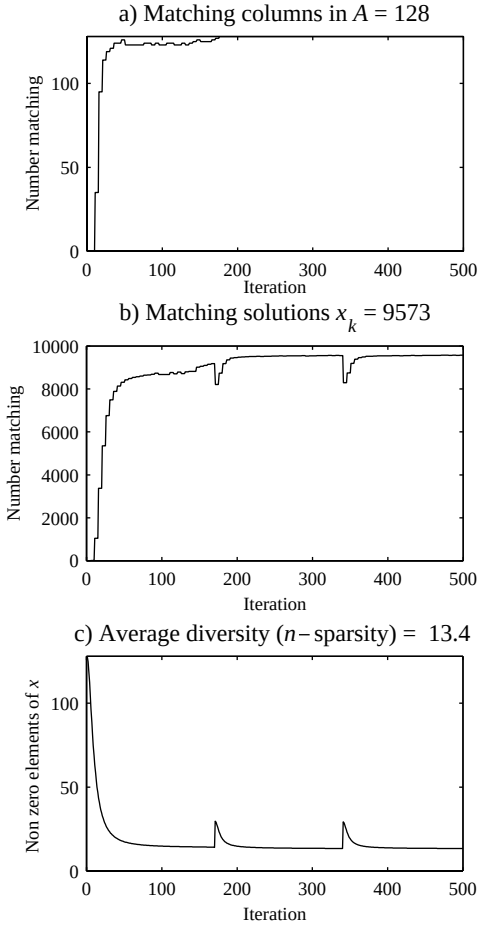

Figure 1: Performance of the learning algorithm with matrix A size 64x128. The spikes in the graphs indicate where some solutions $\widehat{x}_k$ were reinitialized because they were not sparse enough.

matrix and $r = 10..15$, the algorithm is able to recover 99.5% (127.4/128) of the columns of A and 94.6% (9463/10,000) of the solutions x_k to within the tolerance given above. This shows a clear improvement over the original FOCUSS-DL algorithm which only learns 80.3% of the A columns and 40.1% of the solutions x_k. Learning curves for one of the trials of this experiment (Figure 1) show that most of the columns of A are learned quickly within the first 100 iterations, and that the diversity of the solutions drops to the desired level. Figure 1b shows that it takes somewhat longer to correctly learn the x_k, and that reinitialization of the low sparsity solutions (at iterations 175 and 350) helps to learn additional solutions. When $\widehat{A}$ is normalized to $||\widehat{A}||_F = 1$ without column normalization, only 83.6% (107/128) of the columns of A are recovered correctly.

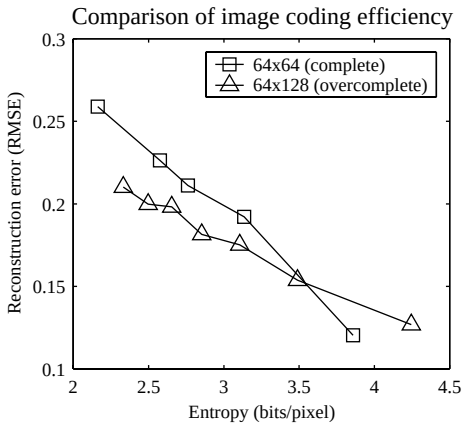

Figure 2: Comparing the coding efficiency of complete and overcomplete representations on 8x8 patches drawn from natural images. The points on the curve are the results from different values of p, at the bottom right, $p = 1.0$, and at the top left, $p = 0.5$. For smaller p, the overcomplete case is more efficient at the same level of reconstruction error (RMSE).

4 Image data experiments

Previous work has shown that learned basis functions can be used to code data more efficiently than traditional Fourier or wavelet bases [7]. The algorithm for finding overcomplete bases in [7] is also designed to solve the problem (1), but differs from our method in a number of ways, including using only the Laplacian prior ($p = 1$), and using conjugate gradient optimization for finding sparse solutions (whereas we use the FOCUSS algorithm). It is widely believed that overcomplete representations are more efficient than complete bases, but in [7] the overcomplete code was less efficient (measured in bits/pixel entropy), and it was suggested that different priors could be used to improve the efficiency. Here, we show that our algorithm is able learn more efficient overcomplete codes for priors with $p < 1$.

The training data consisted of 10,000 8x8 image patches drawn at random from black and white images of natural scenes. The parameter p was varied from $0.5..1.0$, and the algorithm was trained for 150 iterations. The complete matrix (64x64) was compared with the 2x overcomplete matrix (64x128). Other parameters were set: $\gamma = 0.01, \lambda_{\text{max}} = 2 \times 10^{-3}$. The coding efficiency was measured using the entropy (bits/pixel) method described in [7]. Figure 2 plots the entropy vs. reconstruction error (root-mean-square-error, RMSE), and shows that when $p < 0.9$ the entropy is less for the overcomplete representation at the same RMSE.

Studies of the human visual cortex have shown a higher

Table 1: Synthetic data results

			Learned A columns			Learned x		
Algorithm	Size of A	Diversity, r	Avg.	Std. dev.	%	Avg.	Std. dev.	%
FOCUSS-DL	20x30	7	25.3	3.4	84.2%	675.9	141.0	67.6%
FOCUSS-CNDL	20x30	7	28.9	1.6	96.2%	846.8	97.6	84.7%
FOCUSS-CNDL	64x128	7	125.3	2.1	97.9%	9414.0	406.5	94.1%
FOCUSS-CNDL	64x128	5-10	126.3	1.3	98.6%	9505.5	263.8	95.1%
FOCUSS-DL	64x128	10-15	102.8	4.5	80.3%	4009.6	499.6	40.1%
FOCUSS-CNDL	64x128	10-15	127.4	1.3	99.5%	9463.4	330.3	94.6%

degree of overrepresentation of the fovea compared to the other mammals, which suggests an interesting connection between overcomplete representations and visual acuity and recognition abilities [13].

5 Conclusions

We have shown that an improved version of the FOCUSS dictionary learning algorithm is able to more accurately recover sparse solutions to blind linear inverse problems. We have also shown that our algorithm with the proper choice of the prior can learn an overcomplete representation which can encode natural images more efficiently than complete bases learned from data (which in turn are more efficient than standard non-adaptive bases, such as Fourier or wavelet bases).

6 Appendix: Column normalization of $\widehat{A}$

The convergence proof of the algorithm requires that the estimates $\widehat{A}$ lie on the surface $||\widehat{A}||_F = 1$, and we further restrict $\widehat{A}$ to the subset of column normalized matrices where $||a_i|| = ||a_j|| = 1/\sqrt{n}$. We show that this still allows us to move from A_{init} to the solution A by showing that the subset of column normalized matrices is *simply connected*, i.e. there is a continuous path between any two column normalized matrices A, A',

$$A = [a_1, ..., a_n], \quad A' = [a'_1, ..., a'_n]$$
$$||a_i|| = ||a_j|| = 1/\sqrt{n} \implies ||A||_F = ||A'||_F = 1$$

There is a continuous path from A to the matrix $[a'_1, a_2, ..., a_n]$ along the hypersphere of dimension m, $||a_1|| = 1/\sqrt{n}$, which keeps the Frobenius norm unity. Using the same argument, there is a continuous path from $[a'_1, a_2, ..., a_n]$ to $[a'_1, a'_2, ..., a_n]$, and so on to $[a'_1, ..., a'_n] = A'$. A convergence proof using column normalization can be constructed and will be presented elsewhere.

References

[1] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," *Electroencephalography and Clinical Neurophysiology*, vol. 95, no. 4, pp. 231–251, 1995.

[2] R. Vigário and E. Oja, "Independence: A new criterion for the analysis of the electromagnetic fields in the global brain?," *Neural Networks*, vol. 13, pp. 891–907, 2000.

[3] K. Kreutz-Delgado and B. D. Rao, "FOCUSS-based dictionary learning algorithms," in *Proceedings of the SPIE Volume 4119: Wavelet Applications in Signal and Image Processing VIII*, vol. 4119-53, (Bellingham, Washington), SPIE - The International Society for Optical Engineering, July-August 2000.

[4] P. S. Bradley, O. L. Mangasarian, and W. N. Street, "Feature selection via mathematical programming," *INFORMS Journal on Computing*, vol. 10, pp. 209–217, 1998.

[5] V. Cherkassky and F. Mulier, *Learning from Data: Concepts, Theory, and Methods*. New York: Wiley, 1998.

[6] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Sig. Proc.*, vol. 47, pp. 187–200, January 1999.

[7] M. S. Lewicki and B. A. Olshausen, "A probabilistic framework for the adaptation and comparison of image codes," *J. Opt. Soc. Am. A*, vol. 16, pp. 1587–1601, July 1999.

[8] A. Hyvärinen, R. Cristescu, and E. Oja, "A fast algorithm for estimating overcomplete ICA bases for image windows," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN'99)*, pp. 894–899, July 1999. Washington, D.C.

[9] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[10] K. Engan, *Frame based signal representation and compression*. PhD thesis, Stavanger Universty College, Norway, 2000.

[11] K. Kreutz-Delgado and B. D. Rao, "A general approach to sparse basis selection: Majorization, concavity, and affine scaling," Tech. Rep. UCSD-CIE-97-7-1, University of California, San Diego, July 1997. http://cairo.ucsd.edu/∼kreutz.

[12] B. D. Rao and K. Kreutz-Delgado, "Basis selection in the presence of noise," in *Conference Record of the 32rd Asilomar Conference on Signals, Systems and Computers*, pp. 752–756, 1998.

[13] Z. Popovic and J. Sjöstrand, "Resolution, separation of retinal ganglion cells, and cortical magnification in humans," *Vision Research*, vol. 41, pp. 1313–1319, 2001.