FINDING SPARSE REPRESENTATIONS IN MULTIPLE RESPONSE MODELS VIA BAYESIAN LEARNING

David P. Wipf and Bhaskar D. Rao

Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407 USA
e-mail: dwipf@ucsd.edu, brao@ece.ucsd.edu

ABSTRACT

Given a large overcomplete dictionary of basis vectors, our goal is to efficiently represent $L > 1$ signal vectors using coefficient expansions marked by a common sparsity profile. This generalizes the standard sparse representation problem to the case where we have access to multiple responses that were putatively generated by the same small subset of features. Ideally, we would like to uncover the associated sparse generating weights, which can have physical significance in many applications. The generic solution to this problem is combinatorial and therefore we seek approximate procedures. Sparse approximation algorithms tailored to the multiple response domain have typically fallen into two categories: Greedy algorithms such as Matching Pursuit or regularized least-squares methods such as Basis Pursuit and FOCUSS. While these approaches have been extensively analyzed by others, there has been comparably less progress with regard to the development new sparse approximation cost functions and algorithms. Herein, we derive an alternative cost function and associated learning rule based upon a sparse Bayesian learning formulation that improves upon existing methods in many cases.

1. INTRODUCTION

Suppose we are presented with a set of L target signals (or responses) in $\mathbb{R}^N$, $T = [\boldsymbol{t}_{\cdot 1}, \ldots, \boldsymbol{t}_{\cdot L}]$, and an overcomplete dictionary of basis vectors, $\Phi \in \mathbb{R}^{N \times M}$, that are linked by a generative model of the form

$$T = \Phi W + \mathcal{E}, \tag{1}$$

where $W = [\boldsymbol{w}_{\cdot 1}, \ldots, \boldsymbol{w}_{\cdot L}] = [\boldsymbol{w}_{1\cdot}; \ldots; \boldsymbol{w}_{M\cdot}]$ is an unknown weight matrix and $\mathcal{E}$ is noise. Moreover, suppose we have some prior belief that every $\boldsymbol{t}_{\cdot j}$ has been generated by a sparse coefficient expansion, each of which is characterized by a common sparsity profile, i.e., most rows of W have zero norm. Such a situation arises in many diverse applications such as neuromagnetic imaging [8], communications [3], and signal processing [6]. Given T and Φ, the estimation goal in each case is to approximate the row-sparse generating weights.

Given a statistical model for the noise $\mathcal{E}$, we can formulate the likelihood for W. Furthermore, if we assume some sparse prior distribution on W, we may then consider finding the posterior distribution $p(W|T)$, which allows us to assess which basis vectors (i.e., columns of Φ) are important in representing T. For example, given a Gaussian likelihood and an ℓ_0-quasi-norm-based row prior, we could search for the posterior mode via

$$\min_{W} \quad \|T - \Phi W\|_{\mathcal{F}}^2 + \lambda \sum_{i=1}^{M} \mathcal{I}\left[\|\boldsymbol{w}_{i\cdot}\|_2 > 0\right]. \tag{2}$$

Unfortunately, this problem has a combinatorial number of suboptimal local minima, as do many MAP estimation problems involving sparsity-inducing priors (an exception being the Laplacian-based Basis Pursuit formulation, which leads to a linear programming implementation). Rather than embarking on a difficult mode-finding expedition, we instead enlist an alternative Bayesian strategy that is concerned with exploring regions of significant probability mass. Specifically, we posit a standard sparse prior and then adopt a convenient class of variational approximations that allow us to directly track areas of probability mass in the full distribution. Moreover, this particular approximation consistently places its prominent posterior mass on the appropriate region of W-space necessary for sparse recovery. The underlying methodology is based on previous work in [12]. Additional details can be found in [13].

2. ALGORITHM DERIVATION

We must first specify the functional forms for our assumed likelihood and prior. Starting with the former, we postulate $p(T|W)$ to be conditionally Gaussian with noise variance σ^2. Thus, for each $\boldsymbol{t}_{\cdot j}$, $\boldsymbol{w}_{\cdot j}$ pair, we have,

$$p(\boldsymbol{t}_{\cdot j}|\boldsymbol{w}_{\cdot j}) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{t}_{\cdot j} - \Phi \boldsymbol{w}_{\cdot j}\|_2^2\right). \tag{3}$$

This research was supported by an ARCS Foundation scholarship, DiMI grant #22-8376, and Nissan.

Next, we will denote our specific hypothesis (or prior belief) that W is row-sparse as $\mathcal{H}$. To encourage few nonzero rows in W (or equivalently, few selected columns of Φ), we choose the sparsity-inducing Student's t-distribution as the basis of $p(W;\mathcal{H})$. Specifically, we adopt the following form for the i-th row of W:

$$p(\boldsymbol{w}_{i\cdot};\mathcal{H}) = C\left(b + \frac{\|\boldsymbol{w}_{i\cdot}\|_2^2}{2}\right)^{-(a+L/2)}, \qquad (4)$$

where a, b, and C are positive constants. Such a prior favors rows with zero norm (and therefore all zero elements) owing to the sharp peak at zero and heavy tails. The row priors may then be multiplied together to form $p(W;\mathcal{H})$. Given these selections, the resulting joint density $p(W,T;\mathcal{H})$ cannot be easily evaluated to affect sparse recovery and so we consider a reasonable approximation. In choosing an approximate model $p(W,T;\hat{\mathcal{H}})$, we would like to match, where possible, significant regions of probability mass in the true model $p(W,T;\mathcal{H})$. For a given T, one obvious way to do this is to select $\hat{\mathcal{H}}$ by minimizing the sum of the misaligned mass, i.e.,

$$\mathcal{L}(\hat{\mathcal{H}}) \triangleq \int \left| p(W,T;\mathcal{H}) - p(W,T;\hat{\mathcal{H}}) \right| dW. \qquad (5)$$

We must now consider a suitable class of approximations $\hat{\mathcal{H}}$ such that (5) is computable. By extending convexity results from [12], we can form a rigorous lower bound to each row prior via

$$p(\boldsymbol{w}_{i\cdot};\mathcal{H}) \geq p(\boldsymbol{w}_{i\cdot};\hat{\mathcal{H}}) \triangleq \exp\left(-\frac{b}{\gamma_i}\right)\gamma_i^{-a}\mathcal{N}(0,\gamma_i I). \quad (6)$$

Combining each of these approximate row priors, we arrive at the full approximate prior $p(W;\hat{\mathcal{H}}) = \prod_i p(\boldsymbol{w}_{i\cdot};\hat{\mathcal{H}})$, whose form is modulated by a vector of variational parameters $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_M]^T$. At this point, we are positioned to minimize (5) using $\hat{\mathcal{H}}$ selected from the set of variational approximations. Specifically, using the $p(W;\hat{\mathcal{H}})$ derived above, (5) simplifies to

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) &\equiv \int -p(T|W)p(W;\hat{\mathcal{H}})dW \\ &\equiv \sum_{j=1}^{L}\frac{1}{2}\left[\log|\Sigma_t| + \boldsymbol{t}_{\cdot j}^T\Sigma_t^{-1}\boldsymbol{t}_{\cdot j}\right] + \\ &\quad \sum_{i=1}^{M}\left(\frac{b}{\gamma_i} + a\log\gamma_i\right), \qquad (7)\end{aligned}$$

where $\Sigma_t \triangleq \sigma^2 I + \Phi\Gamma\Phi^T$, $\Gamma \triangleq \text{diag}(\boldsymbol{\gamma})$, and the variational assumptions have conveniently allowed us to remove the absolute value and therefore, explicit dependency on $p(W,T;\mathcal{H})$. Treating the unknown weights W as hidden data, we can optimize this expression using a simple EM algorithm. For the E-step, this requires computation of the posterior moments

$$\begin{aligned}\Sigma &\triangleq \text{Cov}[\boldsymbol{w}_{\cdot j}|T;\boldsymbol{\gamma}] = \Gamma - \Gamma\Phi^T\Sigma_t^{-1}\Phi\Gamma, \qquad (8)\\ \mathcal{M} &= [\boldsymbol{\mu}_{\cdot 1},\ldots,\boldsymbol{\mu}_{\cdot L}] \triangleq \text{E}[W|T;\boldsymbol{\gamma}] = \Gamma\Phi^T\Sigma_t^{-1}T,\end{aligned}$$

while the M-step is expressed via the update rule

$$\gamma_i^{(\text{new})} = \frac{\frac{1}{L}\|\boldsymbol{\mu}_{i\cdot}\|_2^2 + \Sigma_{ii} + 2b}{1+2a} \qquad (9)$$

for all $i = 1,\ldots,M$. Interestingly, when $L = 1$ both (8) and (9) can be reduced to the sparse Bayesian learning (SBL) iterations (the EM version) derived in [10] for single response models. Because of this affiliation, we shall refer to this multiple response algorithm as M-SBL, for ***M**ultiple* response model ***S**parse **B**ayesian **L**earning*.

Using the specified update rules for $\boldsymbol{\gamma}$, we arrive at the approximate posterior $p(W|T;\boldsymbol{\gamma}^*)$, where $\boldsymbol{\gamma}^*$ indicates a fixed point of (8) and (9). This distribution is multivariate Gaussian with moments given by (8). Also, we find that when we choose $a = b \to 0$, many of the γ_i^*'s are equal to zero. This effectively collapses $p(W|T;\boldsymbol{\gamma}^*)$ to a relevant low-dimensional affine subspace of W-space. This collapse, and the Gaussian nature of the nonzero portion of the distribution, allows us to easily evaluate the posterior weight mass in assessing the relative importance of each basis vector. Specifically, as we will soon show, if a vector plays an important role in shaping T, then substantial posterior mass will be placed in regions where the corresponding weights are nonzero. In contrast, superfluous basis vectors are effectively pruned by lumping all posterior mass at the appropriate zero-valued weights. Moreover, a common sparsity profile is ensured since $\boldsymbol{\gamma}$ is the same for each column of W, i.e., a single γ_i for each row.

While the algorithm was derived using real quantities for simplicity, it is easily extensible to the complex domain [13]. Additionally, the multiple response SBL framework naturally allows for learning the dictionary Φ (of course now we can no longer assume a common sparsity profile). As a direction for future study, an especially efficient and robust algorithm exists when the unknown dictionary is complete and possibly constrained to being orthogonal. The latter case is useful for learning a sparsity-inducing transform for wavelet shrinkage.

3. EMPIRICAL COMPARISONS

In [2, 3, 7, 9, 11], several methods are presented for solving estimation problems based on (1). These algorithms represent multiple response extensions of more familiar methods such as Orthogonal Matching Pursuit, Basis Pursuit, and FOCUSS, hence we will refer to them as M-OMP, M-BP, and M-FOCUSS respectively. The latter two approaches

can be formulated as MAP estimation using a Gaussian likelihood model and an implicit, ℓ_p-quasi-norm-based prior that penalizes diversity (i.e., rewards sparsity). In fact, M-FOCUSS (with $p \rightarrow 0.0$) implicitly employs an equivalent prior to M-SBL when we choose $a = b \rightarrow 0$; both are related to a multiple response version of the Jeffreys noninformative prior. The primary difference is that during optimization, M-SBL is traversing a restricted space of posterior mass, whereas the others search for the posterior mode.

We would like to quantify the performance of M-SBL relative to these other methods in recovering sparse sets of generating weights, which in many applications have physical significance (e.g., source localization). To accommodate this objective, we performed a series of simulation trials where by design we have access to the sparse, underlying model coefficients. For simplicity, noiseless tests were performed first; this facilitates direct comparisons because discrepancies in results cannot be attributed to poor selection of trade-off parameters (which balance sparsity and quality of fit) in the case of most algorithms.

Each trial consisted of the following: First, an overcomplete $N \times M$ dictionary Φ is created with columns draw uniformly from the surface of a unit hypersphere as proposed in [4]. Sparse weight vectors $\boldsymbol{w}_{\cdot 1}, \ldots, \boldsymbol{w}_{\cdot L}$ are randomly generated with D nonzero entries and a common sparsity profile. Nonzero amplitudes are drawn from a uniform distribution. Response values are then computed as $T = \Phi W$. Each algorithm is presented with T and Φ and attempts to estimate W. With M-OMP, M-BP and M-FOCUSS, this is accomplished directly. In contrast, M-SBL produces a tractable posterior distribution on W, which we may then employ to make a prediction, namely, we choose the posterior mean $\mathcal{M}^*$ as our estimate. For all methods, we can compare W with $\hat{W}$ after each trial to see if the sparse generating weights have been recovered. Results are shown in Figure 1 (*top*) and (*middle*) as L and M are varied.

We also performed analogous tests with the inclusion of noise. Specifically, AWGN was added to produce an SNR of 10dB. When noise is present, we do not expect to reproduce T exactly, so we now classify a trial as successful if the norm of each estimated row associated with a nonzero row of W is greater than the norms of all other rows. Figure 1 (*bottom*) displays sparse recovery results as the trade-off parameter for each algorithm is varied.

Overall, we see that M-SBL outperforms the others. It is especially salient to compare the M-SBL results with M-FOCUSS, $p = 0.0$. Based on equivalent implicit Bayesian priors, we see that optimizing with respect to mass handily outperforms mode-finding. Additionally, Figure 1 is representative of a larger set of trials based on diverse experimental conditions, e.g., different dictionary types, alternate weight distributions, use of complex data, etc. In the cases tested thus far, M-SBL has maintained a higher probability of recovering the sparse generative weights.

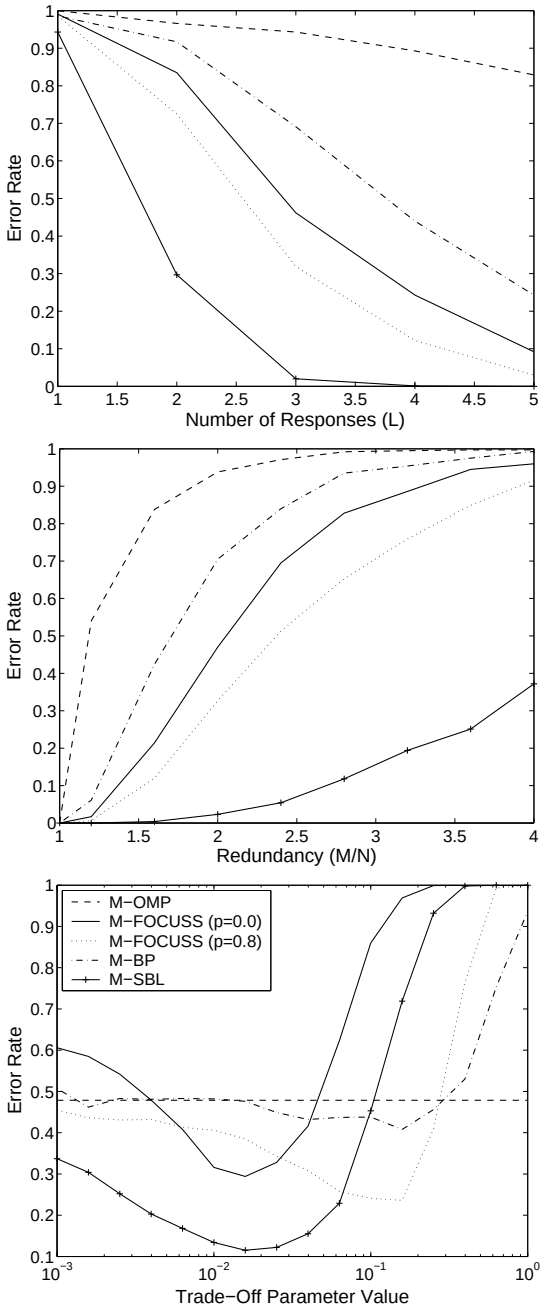

Fig. 1. Empirical results comparing the probability that each algorithm fails to find the sparse generating weights under various testing conditions. Each data point is based on 1000 independent trials, and in all cases $N = 25$. *Top*: With $M = 50$ and $D = 16$, we vary L from 1 to 5 to examine the benefits of multiple responses. *Middle*: With $D = 16$ and $L = 3$, M is varied from N to $4N$ to explore the ability of each algorithm to handle increased dictionary redundancy. *Bottom*: With 10dB additive Gaussian noise, $M = 100$, $D = 8$, and $L = 4$, we vary the trade-off parameter for each algorithm. Note that M-OMP performance is flat here since it has no trade-off parameter.

4. ANALYSIS

Several interesting things are worth noting with respect to the M-SBL framework. First, in the absence of noise, if $\boldsymbol{\gamma}^*$ is a global minimizer of $\mathcal{L}(\boldsymbol{\gamma})$, then $\mathcal{M}^*$ is the maximally row-sparse solution to $T = \Phi W$ under reasonable conditions [13]. Thus, when the algorithm fails, it is because of convergence to a local minimum (as can occur with M-FOCUSS) rather than convergence to a global minimum that is not maximally sparse (as can occur with M-BP). However, it appears that avoidance of direct mode-seeking reduces the number of problematic local minima that exist. For example, we offer the following result [13]:

Result: Given a dictionary Φ with $\text{spark}(\Phi) = N+1$ and a set of responses T, let W_0 be the maximally row-sparse solution to $T = \Phi W$, with $D_0 < N$, $D_0 \leq L$ nonzero rows. Then $\mathcal{L}(\boldsymbol{\gamma})$ has no (non-global) local minima if the nonzero rows of W_0 are orthogonal and $\sigma^2 = 0$.

Consequently, in this restricted setting, M-SBL will always find W_0 unlike M-OMP, M-BP, or M-FOCUSS, all of which may fail under the stipulated conditions (facts that we have verified experimentally). When noise is present, it becomes significantly more difficult to provide any guarantees with regard to local minima avoidance. However, in the special case where $\Phi^T\Phi = I$, it is not difficult to show that no M-SBL local minima exist, unlike M-FOCUSS which can have up to 2^M local minima as $p \rightarrow 0$. Moreover, the unique, globally minimizing stationary point $\boldsymbol{\gamma}^*$ produces a posterior mean of which each row satisfies,

$$\boldsymbol{\mu}_{i\cdot}^* = \boldsymbol{w}_{i\cdot}^{\text{MN}} \left(1 - \frac{L\sigma^2}{\|\boldsymbol{w}_{i\cdot}^{\text{MN}}\|_2^2}\right)^+, \tag{10}$$

where $W^{\text{MN}} \triangleq \Phi^\dagger T$ and $(\cdot)^+$ zeroes negative values. As it turns out, these weight estimates represent a direct, multiple-response extension of those obtained for this problem using the nonnegative garrote estimator [1, 5]. Moreover, we obtain added robustness to noise because the threshold operator is moderated by an average across responses. Consequently, in this setting M-SBL can be interpreted as a form of generalized shrinkage method, truncating small values to zero and shrinking others by a factor that decreases as the magnitude grows (likewise, with an orthonormal dictionary M-BP becomes a generalized soft-threshold estimator).

With regard to computational comparisons, each M-SBL and M-FOCUSS iteration is $O(MN^2 + NML)$ for real or complex data (assuming $N \leq M$). In contrast, the second-order cone (SOC) implementation of M-BP [7] is $O(L^3M^3)$. This could be prohibitively large for $M \gg N$, although fewer total iterations are usually possible. Of course M-OMP is decidedly less costly than all of these methods.

In conclusion, empirical evidence offers some of the strongest support for M-SBL as a viable candidate for sparse recovery tasks in multiple response models. Preliminary results in various source localization applications are very promising.

5. REFERENCES

[1] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[2] J. Chen and X. Huo, "Sparse representations for multiple measurement vectors (MMV) in an overcomplete dictionary," *IEEE Int. Conf. on Accoustics, Speech, and Signal Processing*, vol. 4, pp. 257–260, March 2005.

[3] S.F. Cotter, B.D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[4] D.L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Stanford University Technical Report*, Sept. 2004.

[5] H. Gao, "Wavelet shrinkage denoising using the nonnegative garrote," *Journal of Computational and Graphical Statistics*, vol. 7, no. 4, pp. 469–488, 1998.

[6] B.D. Jeffs, "Sparse inverse solution methods for signal and image processing applications," *IEEE Int. Conf. on Accoustics, Speech, and Signal Processing*, vol. 3, pp. 1885–1888, May 1998.

[7] D.M. Malioutov, M. Çetin, and A.S. Willsky, "Sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[8] J.W. Phillips, R.M. Leahy, and J.C. Mosher, "MEG-based imaging of focal neuronal current sources," *IEEE Trans. Medical Imaging*, vol. 16, no. 3, pp. 338–348, 1997.

[9] B.D. Rao and K. Kreutz-Delgado, "Basis selection in the presence of noise," *Proc. 32nd Asilomar Conf. on Signals, Systems and Computers*, vol. 1, pp. 752–756, Nov. 1998.

[10] M.E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[11] J.A. Tropp, A.C. Gilbert, and M.J. Strauss, Simultaneous sparse approximation via greedy pursuit. *IEEE Int. Conf. on Accoustics, Speech, and Signal Processing*, vol. 5, pp. 721–724, March 2005.

[12] D.P. Wipf, J.A. Palmer, and B.D. Rao, "Perspectives on sparse Bayesian learning," *Advances in Neural Information Processing Systems*, vol. 16, 2004.

[13] D.P. Wipf and B.D. Rao, "A variational Bayesian strategy for solving the simultaneous sparse approximation problem," *UC San Diego Technical Report*, Oct. 2005.