Augmented Bayesian Compressive Sensing

David Wipf[1], Jeong-Min Yun[2], and Qing Ling[3]

[1]Visual Computing Group
Microsoft Research
Beijing, 100080, China
davidwip@microsoft.com

[2]Dept. of Computer Science
POSTECH
Pohang, 790-784, Korea
azida@postech.ac.kr

[3]Dept. of Automation
USTC
Hefei, Anhui, 230027, China
qingling@mail.ustc.edu.cn

Abstract

The simultaneous sparse approximation problem is concerned with recovering a set of multichannel signals that share a common support pattern using incomplete or compressive measurements. Multichannel modifications of greedy algorithms like orthogonal matching pursuit (OMP), as well as convex mixed-norm extensions of the Lasso, have typically been deployed for efficient signal estimation. While accurate recovery is possible under certain circumstances, it has been established that these methods may all fail in regimes where traditional subspace techniques from array processing, notably the MUSIC algorithm, can provably succeed. Against this backdrop several recent hybrid algorithms have been developed that merge a subspace estimation step with OMP-like procedures to obtain superior results, sometimes with theoretical guarantees. In contrast, this paper considers a completely different approach built upon Bayesian compressive sensing. In particular, we demonstrate that minor modifications of standard Bayesian algorithms can naturally obtain the best of both worlds backed with theoretical and empirical support, surpassing the performance of existing hybrid MUSIC and convex simultaneous sparse approximation algorithms, especially when poor RIP conditions render alternative approaches ineffectual.

1 Introduction

Our starting point is the generative model

$$Y = \Phi X_0 + \mathcal{E}, \tag{1}$$

where $\Phi \in \mathbb{R}^{n\times m}$ is a dictionary of basis vectors that we assume to have unit ℓ_2 norm, $X_0 \in \mathbb{R}^{m\times t}$ is a matrix of unknown coefficients we would like to estimate, $Y \in \mathbb{R}^{n\times t}$ is an observed response matrix, and $\mathcal{E}$ is noise. The objective is to estimate the unknown generative X_0 under the assumption that it is row-sparse, meaning that many *rows* of X_0 are equal to zero. The problem is compounded appreciably by the additional assumption that $m > n$, meaning the dictionary Φ is overcomplete. When $t = 1$, this reduces to the canonical sparse estimation of a coefficient vector with mostly zero-valued entries or minimal ℓ_0 norm [5]. In contrast, estimation of X_0 with $t > 1$ represents the more general simultaneous sparse approximation problem [3, 18] relevant in numerous domains such as compressive sensing and multi-task learning [9, 14, 19, 22], manifold learning [16], array processing [13], and functional brain imaging [1].

One possibility for estimating X_0 involves solving

$$\min_X \|Y - \Phi X\|_{\mathcal{F}}^2 + \lambda d(X), \quad \lambda > 0, \quad d(X) \triangleq \sum_{i=1}^{m} \mathcal{I}\left[\|\boldsymbol{x}_{i\cdot}\| > 0\right], \tag{2}$$

where the indicator function $\mathcal{I}\left[\|\boldsymbol{x}\| > 0\right]$ equals one if $\|\boldsymbol{x}\| > 0$ and is zero otherwise (here $\|\boldsymbol{x}\|$ is an arbitrary vector norm). $d(X)$ penalizes the number of rows in X that are not exactly equal to zero; for nonzero rows there is no additional penalty for large magnitudes. Moreover, it reduces to the ℓ_0 norm when $t = 1$, i.e., $d(\boldsymbol{x}) = \|\boldsymbol{x}\|_0$, or a count of the nonzero elements in the vector $\boldsymbol{x}$. Note that to facilitate later analysis, we define $\boldsymbol{x}_{\cdot i}$ as the i-th column of matrix X while $\boldsymbol{x}_{i\cdot}$ represents the i-th row.

For theoretical inquiries, asymptotic regimes with $t \to \infty$, or low-noise environments, it is convenient to consider the limit as $\lambda \to 0$, in which case (2) reduces to

$$\min_X d(X), \qquad \text{s.t. } \Phi X_0 = \Phi X. \tag{3}$$

Unfortunately, solving (3) (or the relaxed version (2)) involves a combinatorial search and is therefore not tractable in practice. This has motivated the broader family of sparse penalized regression problems of the form

$$\min_X \sum_i h\left(\|\boldsymbol{x}_{i\cdot}\|_2\right), \qquad \text{s.t. } \Phi X_0 = \Phi X, \tag{4}$$

where h is a non-decreasing, typically concave function.[1] Common examples include $h(z) = z^p, p \in (0, 1]$ [15] and $h(z) = \log(z + \alpha), \alpha \geq 0$ [2]. The parameters p and α are often heuristically selected on an application-specific basis. In particular, when $h(z) = z$, several theoretical results stipulate conditions whereby we are guaranteed to recover X_0 [6]. We refer to this variant as M-Lasso, for multiple-response vector Lasso. Alternatively, greedy methods such as orthogonal matching pursuit (OMP) have been adapted to handle $t > 1$; this we likewise refer to as M-OMP.

However, curiously there remain important special cases where computing X_0 is relatively easy using conventional subspace techniques originating in the array processing community, and yet solving (4) or greedy pursuit methods are guaranteed to fail in producing X_0. In brief, this failure is a consequence of under-utilizing correlation information in Y and ultimately X_0. We review these special cases linked to the classical MUSIC algorithm [7, 8] in Section 2, as well as hybrid methods that combine a subspace estimation step akin to MUSIC with M-OMP-like procedures to obtain superior results, sometimes with theoretical guarantees. While effective when Φ satisfies standard RIP conditions, we have observed that the performance of these hybrids nonetheless degrades substantially outside of idealized conditions.

Section 3 then describes an alternative strategy for estimating X_0 which builds upon multiple-response model Bayesian compressive sensing (BCS) [9, 20]. Here we motivate an enhanced version of canonical Bayesian algorithms that automatically compensates for correlation structure in X_0 without requiring a separate subspace estimation step as in previous methods. In Section 4 we provide theoretical support suggesting that the resulting modified cost function possesses quantifiable advantages over both M-Lasso and optimal MUSIC. Finally, Section 5 reviews related work, and later numerical experiments in Section 6 corroborate our theoretical findings.

[1] Other row norms, such as the ℓ_∞, have been considered as well but are less prevalent.

For convenience define $k \triangleq d(X_0)$. To ensure estimation of X_0 is feasible we must then have $k < n$. Moreover, we henceforth assume without loss-of-generality that $t \leq k$ and rank$[Y] = t$. This is possible because the success or failure of all algorithms considered herein only depends on $YY^\top$. Therefore if $t > k$ or rank$[Y] < t$, we can equivalently collapse the constraint set to $\tilde{Y} = \Phi X$, where $\tilde{Y} = US^{1/2} \in \mathbb{R}^{n \times \text{rank}[Y]}$ with $USV^\top$ equal to the svd decomposition of Y.

We also will assume that spark$[\Phi] = n + 1$, where matrix *spark* quantifies the smallest number of linearly dependent columns [5]. Consequently, the spark condition is equivalent to saying that each $n \times n$ sub-matrix of Φ is full rank. This assumption is adopted for simplicity and our conclusions generalize to other spark values. However, this is a relatively weak condition anyway that will be satisfied almost surely for any dictionary generated before column normalization as

$$\Phi = A + \epsilon R, \tag{5}$$

where A is an arbitrary matrix, R is a random matrix with iid elements drawn from any continuous distribution, and $\epsilon > 0$ is an arbitrarily small constant.

Now consider solving (3) using the MUSIC algorithm. In the present context this involves estimating the row-support of X_0 as follows. First let U denote any orthonormal basis for range$[Y]$. Therefore $I - UU^T$ is a projection operator onto the orthogonal complement of range$[Y]$. We next compute the set $\Omega = \{j \in 1, \ldots, m : \|(I - UU^\top)\boldsymbol{\phi}_{\cdot j}\|_2 \leq \theta_k\}$, where θ_k is chosen such that $|\Omega| = k$, i.e., we are choosing the indeces of the k smallest values of $(I - UU^\top)\boldsymbol{\phi}_{\cdot j}$. Once the support Ω is computed in this way, X_0 can be estimated using $\hat{X} = \Phi_\Omega^\dagger Y$, where Φ_Ω denotes the columns of Φ indexed by Ω.

Interestingly, in the special case where $t = k$ and given our spark assumption, MUSIC is guaranteed to produce an $\hat{X}$ equal to X_0 [7]. This is because range$[Y]$ = range$[\Phi_\Omega]$ under the stated conditions, and therefore $\|(I - UU^\top)\boldsymbol{\phi}_{\cdot j}\|_2 = 0$ iff $j \in \Omega_0$, the support of X_0. Curiously though, solving (4) (or related M-OMP) do not share the same success.

Lemma 1. There will always exist dictionaries Φ with spark$[\Phi] = n + 1$ and coefficients X_0 with $t = k$, such that the optimization problem (4) with any possible h will have minimizing solutions not equal to X_0.

All proofs will be deferred to a subsequent journal publication because of space considerations. Note that the MUSIC algorithm reduces to a simple thresholding procedure when $t = 1$. In fact, it is formally equivalent to selecting the k basis vectors such that $\boldsymbol{y}^T\boldsymbol{\phi}_{\cdot j}$ is maximized, where $Y = \boldsymbol{y}$ is now a vector. However, as t approaches k its performance nears optimality, exceeding the performance of far more complex iterative optimization procedures.

Ideally then we would like to maximally leverage the best of both subspace techniques like MUSIC and more typical multiple response compressive sensing algorithms such that we obtain optimal performance for all values of t relative to k. Several recent methods attempt just this, combining iterative, M-OMP-like greedy updates with a subspace estimation step [4, 10–12]. While these approaches can be interpreted from different perspectives and have subtle differences, the core concept is to estimate a certain portion of the support,

generally $k-t$ elements, using something related to M-OMP, and then apply MUSIC to an augmented subspace to obtain the final t elements of the support.

Although results appear promising, and the core idea is quite compelling, empirically we have observed that these algorithms can all be sensitive to correlation structure in Φ, and regular M-Lasso can display superior performance in certain regimes. To this end, we motivate a simple alternative Bayesian algorithm that seamlessly merges the strengths of both MUSIC and conventional compressive sensing techniques to achieve state-of-the-art performance with novel theoretical properties.

3 Revisiting Bayesian Compressive Sensing

Bayesian compressive sensing is built upon the probabilistic model from [9, 20]. Here we review the basic elements. We start with the Gaussian likelihood and prior distribution

$$p(Y|X) \propto \exp\left[-\frac{1}{2\lambda}\|Y-\Phi X\|_{\mathcal{F}}^2\right] \quad \text{and} \quad p(X;\Gamma) \propto \exp\left[-\frac{1}{2}\text{tr}\left(X^\top \Gamma^{-1} X\right)\right] \tag{6}$$

respectively, where λ is the noise variance (assumed to be known here) and Γ is a diagonal matrix of hyperparameters controlling the prior variance of each row of X. Given this Γ, the posterior distribution $p(X|Y)$ is a Gaussian with mean

$$\hat{X} = \Gamma\Phi^\top\left(\lambda I + \Phi\Gamma\Phi^\top\right)^{-1} Y. \tag{7}$$

Note that when $\lambda \to 0$, this $\hat{X}$ is feasible, i.e., $Y = \Phi\hat{X}$. The central estimation problem then boils down to determining Γ, which can be accomplished using an empirical Bayesian procedure. The basic idea is to integrate out X and solve

$$\max_{\Gamma \geq 0} \int p(Y|X)p(X;\Gamma)dX. \tag{8}$$

Once some Γ^* is computed in this way, we can plug this value into (7) for our estimate of X_0. If diagonal elements of this Γ^* are zero, then the corresponding rows of $\hat{X}$ are also necessarily zero.

Returning to connections with subspace methods, let $\bar{X}_0 \in \mathbb{R}^{k\times t}$ denote the nonzero rows of some X_0, which is necessarily full-rank by our previous assumptions. It has already been demonstrated that if $\bar{X}_0$ has orthogonal rows, then (8) in the limit as $\lambda \to 0$ has a single stationary point Γ^* with sparsity profile matching the true X_0 [20]. But in general we may expect that $\bar{X}_0\bar{X}_0^\top$ potentially possesses significant off-diagonal structure. The most natural way to compensate for such structure within the context of this Bayesian model is to reparameterize the likelihood function as

$$p(Y|X;\Psi) \propto \exp\left[-\frac{1}{2\lambda}\|Y-\Phi X\Psi\|_{\mathcal{F}}^2\right], \tag{9}$$

where Ψ represents an unknown matrix that accounts for coefficient structure when we view $Z_0 \triangleq X_0\Psi$ as the new coefficient set that we wish to recover. We must now jointly estimate Γ and Ψ by maximizing the modified form of (8) with $p(Y|X;\Psi)$ replacing $p(Y|X)$.

Retracing back to our original objective of solving (3), then in the limit as $\lambda \rightarrow 0$, maximizing $\int p(Y|X;\Psi)p(X;\Gamma)dX$ is equivalent to solving

$$\lim_{\lambda \rightarrow 0} \min_{\Psi;\Gamma \succeq 0} -2 \log \int p(Y|X;\Psi)p(X;\Gamma)dX \quad \equiv \quad \lim_{\lambda \rightarrow 0} \min_{\Psi;\Gamma \succeq 0} L(\Gamma,\Psi). \tag{10}$$

The relevant cost function is defined as

$$L(\Gamma,\Psi) \triangleq \text{tr}\left[Y\left(\Psi^{\top}\Psi\right)^{-1}Y^{\top}\left(\Phi\Gamma\Phi^{\top}+\lambda I\right)^{-1}\right] + t\log\left|\Phi\Gamma\Phi^{\top}+\lambda I\right| + n\log\left|\Psi^{\top}\Psi\right| \tag{11}$$

after computing the required integral and discarding irrelevant constants. Additionally, the limit must be taken outside of the minimization in (10). The reason we consider this limit rather than simply $\lambda = 0$ is for technical reasons related to the situation where $\Phi\Gamma\Phi^{\top}$ is no longer full rank.

If we optimize first over Ψ, there exists a closed-form solution such that

$$\Psi^{\top}\Psi = \frac{1}{n}Y^{\top}\left(\Phi\Gamma\Phi^{\top}+\lambda I\right)^{-1}Y. \tag{12}$$

This can be determined by computing gradients, equating to zero, and checking the requisite optimality conditions. Plugging this expression into (11) we obtain the reduced equivalent cost function

$$\mathcal{L}(\Gamma) \triangleq t\log\left|\Phi\Gamma\Phi^{\top}+\lambda I\right| + n\log\left|Y^{\top}\left(\Phi\Gamma\Phi^{\top}+\lambda I\right)^{-1}Y\right|, \tag{13}$$

where parameter-independent terms have been removed. As we will see in Section 4, this penalty function has several desirable attributes relevant to addressing (3) when $\lambda \rightarrow 0$. While not our focus here, noisy variants can also prove effective, although theoretical study is much more difficult because of the intrinsic non-convexity of the objective.

From a practical standpoint, a family of iterative reweighting procedures can be applied to solve $\min_{\Gamma \succeq 0} \mathcal{L}(\Gamma)$ for any value of λ. Arguably the simplest is a form of iterative reweighted least squares (IRLS) based upon a standard majorization-minimization scheme frequently used in BCS. This procedure can exploit only partially solving (11) for fixed Ψ at each iteration. The complete algorithm is as follows. Initialize $\Psi = I$ and $\Gamma = I$. Then compute (7), $\hat{Z} = \hat{X}\Psi$, and

$$S \triangleq \Gamma - \Gamma\Phi^{\top}\left(\lambda I + \Phi\Gamma\Phi^{\top}\right)^{-1}\Phi\Gamma. \tag{14}$$

We then update Γ using

$$\Gamma_{ii} = \frac{1}{t}\sum_{j=1}^{t}\hat{z}_{ij}^2 + s_{ii}. \tag{15}$$

Finally Ψ is updated via (12) and the process is repeated until convergence. These iterations are guaranteed to reduce or leave unchanged (13) at every iteration and can even handle $\lambda \rightarrow 0$ with appropriate use of the pseudo-inverse. However, provable convergence to a stationary point remains an outstanding issue. In this regard iterative reweighted M-Lasso implementations are more amenable to analysis, but we defer this to a subsequent journal article. Regardless of implementational specifics, it is largely the nature of the underlying Bayesian-inspired cost function that contributes to effective deployment. We will consider such issues both theoretically (Section 4) and empirically (Section 6).

4 Analysis

The cost function $\mathcal{L}(\Gamma)$, and its predecessor in (11), are non-convex and seemingly difficult to untangle. However, certain intrinsic properties make them notably appropriate for solving (3), and in particular, speak to an intimate connection with the MUSIC algorithm.

Theorem 1. Assume spark$[\Phi] = n + 1$ and that X_0 is a unique, optimal solution to (3). Then if rank$[X_0] = k = t$, the problem $\lim_{\lambda \to 0} \min_{\Gamma \succeq 0} \mathcal{L}(\Gamma)$ has a single stationary point Γ^*, and this point satisfies $\Gamma^* \Phi^\top \left(\Phi \Gamma^* \Phi^\top\right)^\dagger Y = X_0$.

This result implies that, in terms of successful recovery, we can enjoy performance at least as good as the MUSIC algorithm operating in its optimal regime of $k = t$ (assuming a convergent algorithm). Moreover, this occurs in a completely integrated fashion unlike previous algorithms. And even when $t < k$, we nonetheless reap benefits of this integration because it can be shown that the underlying augmented BCS cost function behaves as though there are $t - 1$ fewer support elements to estimate.

Additionally, if we minimize $\mathcal{L}(\Gamma)$ using an iterative reweighted M-Lasso algorithm with guaranteed convergence to a stationary point, then it can be shown that we will never do worse that M-Lasso either. Hence MUSIC and M-Lasso performance can in some sense be interpreted collectively as lower bounds on the performance of augmented BCS.

However, this alone does not ensure that there exist regimes where a provable performance gain can be expected. In contrast, the following result related to the cost function from (11) can be applied towards this purpose:

Theorem 2. Assume that Φ satisfies (5) and let $\pi_{(i)}[X]$ denote the value of the i-th largest ℓ_2 row-norm of a matrix X. Then there exists a set of $n - 2$ constants $\nu_i \in (0, 1]$ such that, for *any* $Y = \Phi X_0 \Psi$ generated with $d(X_0) < n$, Ψ invertible, and

$$\pi_{(i+1)}[X_0 \Psi] \leq \nu_i \pi_{(i)}[X_0 \Psi], \quad i = 1, \ldots, n-2, \tag{16}$$

the following two conditions will always hold:

(I) The problem $\lim_{\lambda \to 0} \min_{\Gamma \succeq 0} L(\Gamma, \Psi)$ has a single stationary point Γ^*, and this point satisfies $\Gamma^* \Phi^\top \left(\Phi \Gamma \Phi^\top\right)^\dagger Y = X_0 \Psi$.

(II) X_0 will be the unique solution to (3).

This result actually applies to the original BCS cost function, which is what (11) reduces to when $\Psi = I$. Hence if we initialize with $\Psi = I$ and the conditions of Theorem 2 hold, and then we optimize only Γ until convergence with a globally convergent algorithm, we are guaranteed to learn an optimal X_0 without ever needing to incorporate any Ψ updates to conceivably avoid local minima. The advantage then of the augmented objective function with general Ψ is that even if Theorem 2 does not hold with $\Psi = I$, it may hold at a later iteration after $\Psi \neq I$ has been updated using (12), allowing remaining local minima to potentially be avoided.

Importantly, Theorem 2 holds even when Φ exhibits arbitrarily strong correlation patterns (by virtue of the influence of A in (5)) and RIP conditions required by existing algorithms do not apply. Moreover, neither M-Lasso nor MUSIC, nor any combination rule which selects the better of the two, can achieve something similar: there will always exist

dictionaries Φ and coefficient matrices X_0, consistent with the stipulations of Theorem 2 such that failure is inevitable, including the special case where $\Psi = I$. Of course other hybrid algorithms could be pieced together using MUSIC and different penalty function selections for h in (4). But it is completely unclear how to design attendant update rules to guarantee anything similar to the augmented Bayesian strategy discussed herein.

This leaves the family of greedy hybrid algorithms proposed in [4, 10–12] for merging with subspace methods. The drawback with these strategies however is twofold. First, as a baseline sparse estimation procedure, solving (4) is generally more powerful than greedy approaches like M-OMP, especially when the former is implemented with convex iterative reweighting procedures. Secondly, existing hybrid subspace algorithms, which update the support in two separate steps, do not fully consider both the effective subspace of Y *and* intermediate coefficient estimates all in an entirely integrated fashion as with $\mathcal{L}(\Gamma)$.

5 Related Bayesian Analytical and Algorithmic Work

A result related to Theorem 2 has been demonstrated in the special case where $t = 1$ [21]. However, this scenario is decidedly much simpler because it can be shown that any local minimum of $\mathcal{L}(\Gamma)$ or $L(\Gamma, \Psi)$ can be achieved with $d(\Gamma) \leq nt$. Therefore when $t = 1$, this implies that we only need consider candidate local minimizers Γ associated with basic feasible solutions, meaning solutions involving at most n columns of Φ making the corresponding sub-matrix of $\Phi^\top \Phi$ invertible given the implicit spark condition. It then follows that relevant terms at each candidate local minima conveniently decouple, greatly simplifying the analysis.

In contrast, with $t > 1$ we have no such luxury because it is not possible to rule out local minimizers with $d(\Gamma) > n$, and hence we are forced to accommodate this more challenging scenario via a different strategy. It is also important to emphasize that just because the row-norm scaling condition of Theorem 2 is satisfied does not imply that it will additionally be satisfied when applied to each column individually. Therefore we cannot simply adopt the original result from [21] in a column-wise fashion to reproduce Theorem 2.

Finally, from an algorithmic standpoint, [23] considers similar modifications of Bayesian compressive sensing intended to address correlations in the rows of X. However, no theoretical justification is provided beyond what is already known for the standard BCS framework. Moreover, there is no discussion of the intimate connection with subspace methods and the MUSIC algorithm.

6 Numerical Validation

Here we briefly describe some simulations that complement our previous analytical findings. In [4, 10–12] a series of experiments are presented that demonstrate the efficacy of hybrid subspace methods. However, the experimental conditions are not necessarily challenging in the sense that for all cases Φ is generated with iid Gaussian elements. In contrast, for the experiments in this section we generate $\Phi = \sum_{i=1}^{n} i^{-1} \boldsymbol{a}_i \boldsymbol{b}_i^\top$, where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are iid standardized Gaussian vectors of appropriate length, and then normalize each column of the resulting dictionary. This selection ensures that Φ exhibits non-trivial correlations

among columns because of the i^{-1} scale factor.

Next we generate nonzero rows of X_0 as $\bar{X}_0 = \sum_{i=1}^{n} i^{-1} \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are again iid Gaussian. This implies that $\bar{X}_0 \bar{X}_0^\top$ should have significant off-diagonal elements, which should favor subspace-based methods over conventional algorithms like M-Lasso. We fix $m = 200$, $k = 20$, and $t \in \{4, 8, 12, 16\}$. For each value of t, we vary n from $k + 1$ to 100, noting that $k + 1$ is the minimum number of measurements such that recovery of X_0 is even theoretically possible.

For evaluation purposes we compare augmented BCS (or ABCS) with regular BCS implemented using IRLS, M-Lasso, and two hybrid compressive sensing MUSIC algorithms. While in reality these algorithms constitute a family with many potential variations, we choose two variants endorsed by the authors of [10–12]. Specifically, we compare with CS-MUSIC, where code was provided by the authors of [12], and sequential CS-MUSIC, with code from the authors of [10, 11]. Both algorithms were given access to the true value of k in all experiments. M-Lasso, BCS, and ABCS do not use prior information regarding k.

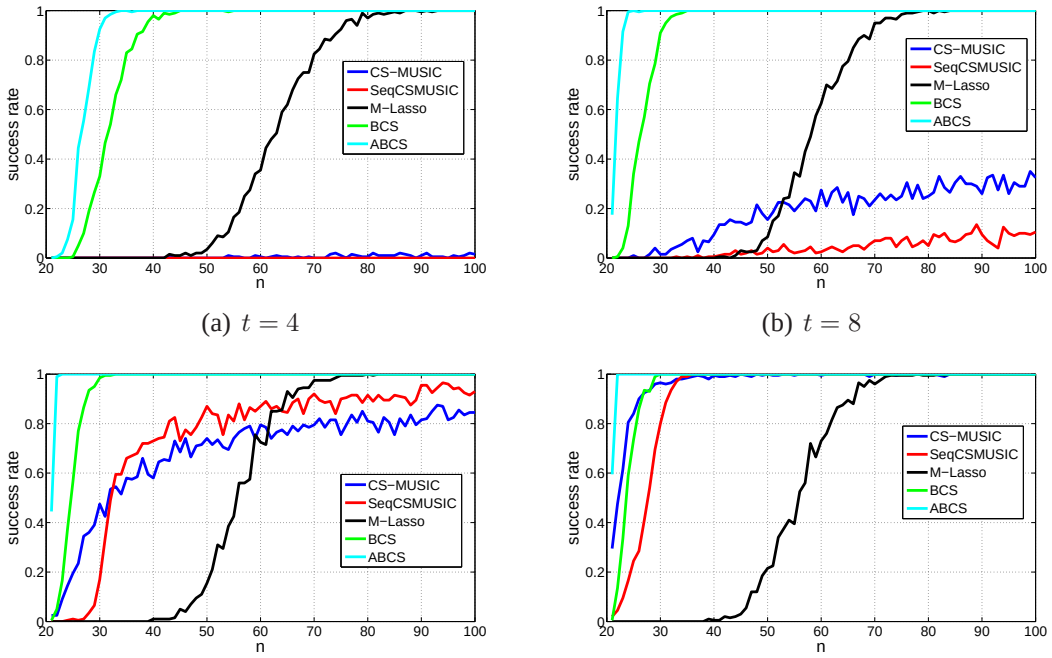
Figure 1 displays the results averaged across 200 independent trials, where the evaluation metric is the frequency of trials where each respective algorithm detects the correct support of X_0. In panel (a) we have the fewest number of snapshots ($t = 4$), and therefore the conditions are least favorable for the hybrid subspace methods. Consequently M-Lasso substantially outperforms both CS-MUSIC and sequential CS-MUSIC. In contrast, as t increases from panels (a) through (b), the subspace approaches acquire additional information such that they eventually can outperform M-Lasso decisively. The latter has difficulty capitalizing on this additional ill-conditioned subspace information and hence M-Lasso displays only marginal improvement from $t = 4$ to $t = 16$.

Regarding ABCS and BCS, they both exhibit excellent performance across all values of t since they have an intrinsic mechanism for compensating for correlations in Φ. However, clearly ABCS is able to more thoroughly exploit subspace information and correlations in $\bar{X}_0$ outperforming the other algorithms in all of the testing conditions.

We next conduct a similar experiment, except now we fix $m = 200$, $t = 10$, and $k \in \{30, 50\}$. We then vary n from $k+1$ to $k+100$. Results are reported in Figure 2. Here we observe that while BCS and ABCS performance is quite stable, the subspace methods, and to some extent M-Lasso, degrade as k and n jointly become larger.

7 Conclusions

Since the original proposal of sparse Bayesian learning algorithms [17], mounting evidence has established that empirical Bayesian techniques can be highly effective for solving sparse linear inverse problems in the $t = 1$ case. As we move to more diverse and structured environments, including the models with row-sparsity considered here, the efficacy of sparse Bayesian estensions has not been fully understood. This work elucidates basic behaviors of BCS and its connection with subspace methods, motivating a targeted enhancement. Although not our focus here, further modifications to accommodate noisy environments can be incorporated using developments from [23], including the ability to estimate the noise level λ automatically.



(a) $t = 4$ (b) $t = 8$

(c) $t = 12$ (d) $t = 16$

Figure 1: Support recovery success rates as the number of measurements n is varied. Each curve represents the average across 200 independent trials. In all cases we fix $m = 200$ and $k = 20$.

References

[1] S. Baillet, J. Mosher, and R. Leahy, "Electromagnetic brain mapping," *IEEE Signal Processing Magazine*, vol. 18, no. 6, pp. 14–30, Nov. 2001.

[2] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[3] S. Cotter, B. 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, April 2005.

[4] M. Davies and Y. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Info. Theory*, vol. 58, no. 2, pp. 1135–1146, Feb. 2012.

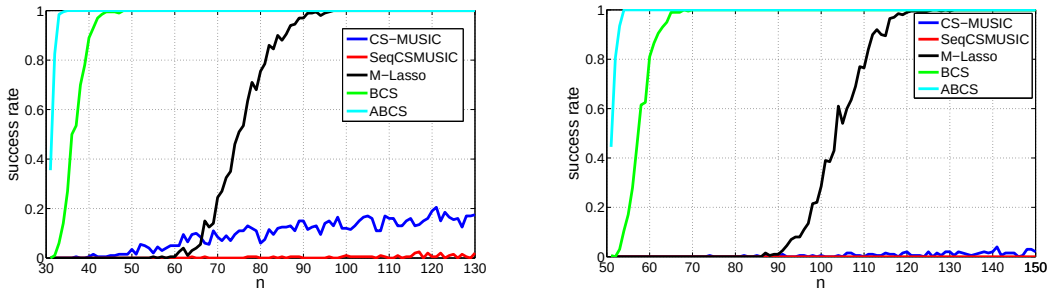
[5] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, March 2003.

[6] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Info. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.

[7] P. Feng, "Universal minimum-rate sampling and spectrum-blind reconstruction for multiband signals," PhD Thesis, University of Illinois, Urbana-Champaign, 1998.

[8] P. Feng and Y. Bresler, "Reduced complexity decision feedback equalization for multipath channels with large delay spreads," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1688–1691, 1996.

[9] S. Ji, D. Dunson, and L. Carin, "Multi-task compressive sensing," *IEEE Trans. Signal Processing*, vol. 57, no. 1, pp. 92–106, Jan. 2009.



(a) $k = 30$ (b) $k = 50$

Figure 2: Support recovery success rates as the number of measurements n is varied. Each curve represents the average across 200 independent trials. In all cases we fix $m = 200$ and $t = 10$.

[10] J. Kim, O. Lee, and J. Ye, "Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing," *IEEE Trans. Info. Theory*, vol. 58, no. 1, pp. 278–301, Jan. 2012.

[11] ——, "Noise robustness in subspace-based joint sparse recovery," *IEEE Trans. Signal Processing*, vol. 60, no. 11, pp. 5799–5809, Nov. 2012.

[12] K. Lee, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Trans. Info. Theory*, vol. 58, no. 6, pp. 3613–3641, June 2012.

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

[14] Q. Ling, Z. Wen, and W. Yin, "Decentralized jointly sparse optimization by reweighted ℓ_q minimization," *IEEE Trans. Signal Processing*, vol. 61, no. 5, pp. 1165–1170, March 2013.

[15] B. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Processing*, vol. 51, no. 3, pp. 760–770, March 2003.

[16] J. Silva, J. Marques, and J. Lemos, "Selecting landmark points for sparse manifold learning," *Advances in Neural Information Processing Systems 18*, pp. 1241–1248, 2006.

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

[18] J. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Processing*, vol. 86, pp. 589–602, April 2006.

[19] M. Wakin, M. Duarte, S. Sarvotham, D. Baron, and R. Baraniuk, "Recovery of jointly sparse signals from a few random projections," *Advances in Neural Information Processing Systems 18*, pp. 1433–1440, 2006.

[20] D. Wipf and B. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Processing*, vol. 55, no. 7, pp. 3704–3716, July 2007.

[21] D. Wipf, B. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Info. Theory*, vol. 57, no. 9, Sep. 2011.

[22] F. Zeng, C. Li and Z. Tian, "Distributed compressive spectrum sensing in cooperative multihop cognitive networks," *IEEE J. Selected Topics in Signal Processing*, vol. 5, no. 2, pp. 37–48, Feb. 2011.

[23] Z. Zhang and B. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE J. Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912–926, Sep. 2011.