Sparse Bayesian Learning via Stepwise Regression: Supplementary Materials

Sebastian Ament [1] **Carla Gomes** [1]

1. Overview

Statements and Proofs We provide the proofs of the results in the paper, including new technical results which might find application in the analysis of related algorithms. In particular, Section A first reviews past results on the optimality of the backward algorithm and proceeds to present technical results that are necessary for the remainder. Section B contains the statements and proofs regarding SBL and RMP and Section C contains the main results and proofs for Stepwise Regression.

Code and Experiments In Section D, we add information about our experiments, include another numerical study of the theoretical results, and show how to incorporate the estimation of the noise variance into RMP, one of the advantages of the probabilistic framework over traditional compressed sensing and feature selection technologies. We highlight the `CompressedSensing.jl` package, which contains our implementations of all algorithms and we hope will serve as a platform for sparsity-inducing algorithms, and the IMCL2021 folder, which contains the experimental setup and results. Associated with each experiment is also an H5-file which holds all the data that is necessary to verify our results, generate the plots of the paper, and rerun the experiments.

[1]Department of Computer Science, Cornell University, Ithaca, NY, USA. Correspondence to: Sebastian Ament <ament@cs.cornell.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

A. Preliminaries

In the following, $\mathbf{\Phi}$ denotes a matrix, and $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$, where $\mathbf{x}$ is a vector with the appropriate dimension. We will refer to $\boldsymbol{\epsilon}$ as a perturbation, and assume the input of the greedy algorithms is $\mathbf{y} + \boldsymbol{\epsilon}$. $\|\cdot\|$ denotes the norm of a linear space, and $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ is the associated metric. We define the distance of a vector $\mathbf{x}$ to a subset $\mathcal{S}$ as

$$d(\mathbf{x}, \mathcal{S}) \stackrel{\text{def}}{=} \min_{\mathbf{s}\in\mathcal{S}} d(\mathbf{x}, \mathbf{s}).$$

$\mathbf{P}_{\mathcal{S}}$ is the least-squares projection onto a subspace $\mathcal{S}$. If $\mathbf{\Phi}$ is a basis for $\mathcal{S}$, $\mathbf{P}_{\mathcal{S}} = \mathbf{\Phi}\mathbf{\Phi}^{+}$. For ease of notation, we define $\mathbf{P}_{\mathcal{A}} = \mathbf{P}_{\text{col}(\mathbf{\Phi}_{\mathcal{A}})}$ for a set of indices $\mathcal{A}$, where $\text{col}(\mathbf{\Phi})$ is the column space of $\mathbf{\Phi}$. Given a subset $\mathcal{A}$ of columns of $\mathbf{\Phi}$, the associated least-squares residual is $\mathbf{r}_{\mathcal{A}} = (\mathbf{I} - \mathbf{P}_{\mathcal{A}})(\mathbf{y} + \boldsymbol{\epsilon})$. For any vector $\mathbf{x}$, $\hat{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|$.

A.1. Prior Work on Backward Elimination

We begin by reviewing the most relevant results on the optimality of the backward algorithm in (Couvreur and Bresler, 2000). First, we define the bisector of two subspaces as the set of equidistant vectors, which is central to the existing theory.

Definition A.1 (Bisector)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two subspaces of a linear space $\mathcal{L}$. Their bisector is*

$$\mathcal{H}(\mathcal{A}, \mathcal{B}) \stackrel{def}{=} \{\mathbf{x} \in \mathcal{L} | d(\mathbf{x}, \mathcal{A}) = d(\mathbf{x}, \mathcal{B})\}.$$

To facilitate notation, we define

$$\mathcal{H}_{ij}(\mathbf{\Phi}) \stackrel{\text{def}}{=} \mathcal{H}(\text{col}(\mathbf{\Phi}_{\setminus i}), \text{col}(\mathbf{\Phi}_{\setminus j})).$$

Using the bisector, (Couvreur and Bresler, 2000) proved the following one-step guarantee for the backward algorithm. Essentially, the proof views the bisectors $\mathcal{H}_{ij}(\mathbf{\Phi})$ as decision boundaries of the algorithm. If the boundaries cannot be crossed due to the magnitude of the perturbation, the algorithm is guaranteed to succeed.

Lemma A.2 ((Couvreur and Bresler, 2000))**.** *Let $\mathbf{x}$ be a k-sparse vector with support set $\mathcal{S}$, and $\mathcal{A} \supset \mathcal{S}$,*

$$\delta = \min_{i\in\mathcal{S}, j\notin\mathcal{S}} d(\mathbf{y}, \mathcal{H}_{ij}(\mathbf{\Phi}_{\mathcal{A}})),$$

and suppose $\|\boldsymbol{\epsilon}\|_2 < \delta$. Then the backward greedy algorithm successfully eliminates an feature that is not in the true support set in the following iteration. That is,

$$\arg\min_{i\in\mathcal{A}} \|\mathbf{r}_{\mathcal{A}\setminus i}\| \notin \mathcal{S}.$$

Proof. The proof is due to (Couvreur and Bresler, 2000). □

Using Lemma A.2, as a building block, the authors provided the following main result on the backward algorithm.

Theorem A.3 ((Couvreur and Bresler, 2000))**.** *Let $\mathbf{x}$ be a k-sparse vector with support set $\mathcal{S}$,*

$$\delta = \min_{k<r\leq n} \min_{|\mathcal{A}|=r} \min_{i\in\mathcal{S}, j\notin\mathcal{S}} d(\mathbf{y}, \mathcal{H}_{ij}(\mathbf{\Phi}_{\mathcal{A}})), \tag{A.1}$$

and suppose $\|\boldsymbol{\epsilon}\|_2 < \delta$. Then the backward greedy algorithm selects the correct support set $\mathcal{S}$ of $\mathbf{x}$ in $m - k$ iterations and the estimate $\mathbf{x}_* = \mathbf{\Phi}_{\mathcal{S}}^{+}\mathbf{y}$ satisfies $\|\mathbf{x}_* - \mathbf{x}\| \leq \delta/\sigma_{\min}(\mathbf{\Phi}_{\mathcal{S}})$.

Proof. See Theorem 1 in (Couvreur and Bresler, 2000). □

Theorem A.3 assumes that there is an underlying sparse vector to be recovered, which corresponds to the assumptions of the sparse recovery problem. The sparse approximation or the subset selection problem does not assume such an underlying sparse vector, but concerns the best possible approximation of a generally non-sparse vector by a sparse vector. Guarantees for the latter are generally weaker. For example, there are approximation guarantees but no conditional optimality guarantees for Forward Selection for the subset selection problem. The following corollary of the previous theorem provides such an optimality guarantee for Backward Elimination.

Corollary A.4 ((Couvreur and Bresler, 2000)). *Let $\mathbf{x}_k$ be the solution to the subset selection problem with sparsity k. That is, $\mathbf{x}_k$ is the k-sparse vector that whose associated residual $\mathbf{r}_k \stackrel{def}{=} \mathbf{y} - \mathbf{\Phi}\mathbf{x}_k$ has the minimum norm among all vectors with at most k non-zero elements. If $\mathbf{r}_k$ satisfies the bound in Theorem A.3 in place of ϵ, the backward algorithm solves the subset selection problem to optimality.*

Proof. See Corollary 2 in (Couvreur and Bresler, 2000). □

However, the authors posit that the expression in equation (A.1) is NP-hard to compute and thus cannot guide a practitioner and confirm whether or not the result of the algorithm is optimal in practice.

A.2. Preliminary Linear Algebraic Results

In the following sections, we present technical results that are necessary to prove the main results, and might be of independent interest for the analysis of related algorithms. We first provide linear algebraic results, followed by two subsections with preliminary results targeted at the forward and backward algorithms, respectively, and finish the section with probabilistic results for Gaussian noise.

Lemma A.5 ((Tropp, 2004)). *Let $\mathbf{\Phi}$ have l_2-normalized columns. The squared singular values σ^2 of a submatrix of $\mathbf{\Phi}$ with at most k columns satisfy $|1 - \sigma^2| \leq \mu_1(k-1)$, where μ_1 is the Babel function of $\mathbf{\Phi}$.*

Proof. This is essentially due to (Tropp, 2004), Lemma 2.3. □

Lemma A.6. *Let $\mathcal{A}$ be a set of column indices of a matrix $\mathbf{\Phi}$, and $\mathcal{A}^c$ be the complement of $\mathcal{A}$ in the set of all column indices of $\mathbf{\Phi}$, and $\mathbf{P}_{\mathcal{A}^c} = \mathbf{\Phi}_{\mathcal{A}^c}\mathbf{\Phi}^{+}_{\mathcal{A}^c}$. Then the eigenvalues of $\mathbf{\Phi}^*_{\mathcal{A}}(\mathbf{I} - \mathbf{P}_{\mathcal{A}^c})\mathbf{\Phi}_{\mathcal{A}}$ are bounded above and below by the minimum and maximum eigenvalue of $\mathbf{\Phi}^*\mathbf{\Phi}$, respectively.*

Proof. See also Lemma 5 in (Cai and Wang, 2011). Note that up to permutation

$$\mathbf{\Phi}^*\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}^*_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}} & \mathbf{\Phi}^*_{\mathcal{A}}\mathbf{\Phi}_{\mathcal{A}^c} \\ \mathbf{\Phi}^*_{\mathcal{A}^c}\mathbf{\Phi}_{\mathcal{A}} & \mathbf{\Phi}^*_{\mathcal{A}^c}\mathbf{\Phi}_{\mathcal{A}^c}. \end{bmatrix}$$

According to a standard result on the block matrix inverse, the lower right block of $(\mathbf{\Phi}^*\mathbf{\Phi})^{-1}$ is $\mathbf{\Phi}^*_{\mathcal{A}}(\mathbf{I} - \mathbf{P}_{\mathcal{A}^c})\mathbf{\Phi}_{\mathcal{A}}$. Therefore, the eigenvalues of the latter matrix are bounded above and below by the extremal eigenvalues of $(\mathbf{\Phi}^*\mathbf{\Phi})^{-1}$. □

Lemma A.7. *Given a matrix $\mathbf{\Phi}$ with columns $\{\boldsymbol{\varphi}_i\}_i$, and an index set $\mathcal{A}$ of size k, let $\boldsymbol{\psi}_i \stackrel{def}{=} (\mathbf{I} - \mathbf{P}_{\mathcal{A}})\boldsymbol{\varphi}_i$ and $i \notin \mathcal{A}$. Then*

$$\sigma_{\min}(\mathbf{\Phi}_{\mathcal{A}\cup i}) \leq \|\boldsymbol{\psi}_i\|_2 \leq \|\boldsymbol{\varphi}_i\|_2.$$

Proof. Since $i \notin \mathcal{A}$, we get using Lemma A.6,

$$\|\boldsymbol{\psi}_i\|_2^2 = \|(\mathbf{I} - \mathbf{P}_{\mathcal{A}})\boldsymbol{\varphi}_i\|_2^2 = \boldsymbol{\varphi}_i^*(\mathbf{I} - \mathbf{P}_{\mathcal{A}})\boldsymbol{\varphi}_i \geq \sigma^2_{\min}(\mathbf{\Phi}_{\mathcal{A}\cup i}),$$

since $\{i\}$ is the complement of $\mathcal{A}$ in $\mathcal{A}\cup\{i\}$. The upper bound is a consequence of $(\mathbf{I} - \mathbf{P}_{\mathcal{A}})$ being a projection, so that $\|\boldsymbol{\psi}_i\| = \|(\mathbf{I} - \mathbf{P}_{\mathcal{A}})\boldsymbol{\varphi}_i\|_2 \leq \|\boldsymbol{\varphi}_i\|_2$. □

Lemma A.8. *Let $\mathcal{A}$ be a set of k indices into $\mathbf{\Phi}$'s columns and $\boldsymbol{\psi}_i = (\mathbf{I} - \mathbf{P}_{\mathcal{A}})\boldsymbol{\varphi}_i$. Then for $i, j \notin \mathcal{A}$ and $i \neq j$,*

$$|\langle \hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_j \rangle| \leq 1 - \sigma_{\min}\left(\mathbf{\Phi}_{\mathcal{A}\cup\{i,j\}}\right)^2.$$

Proof. Define $\mathbf{M} = \mathbf{\Phi}^*_{\{i,j\}}(\mathbf{I} - \mathbf{P}_{\mathcal{A}})\mathbf{\Phi}_{\{i,j\}}$, $\mathbf{D} = \operatorname{diag}(\|\boldsymbol{\psi}_i\|, \|\boldsymbol{\psi}_j\|)$, and $\tilde{\mathbf{M}} = \mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1}$. Then by definition of the determinant, and the determinant of a matrix being the product of its eigenvalues,

$$\begin{aligned} 1 - |\langle \hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_j \rangle|^2 &= \det\left(\tilde{\mathbf{M}}\right) \\ &= \lambda_{\min}\left(\tilde{\mathbf{M}}\right)\lambda_{\max}\left(\tilde{\mathbf{M}}\right) \\ &= \lambda_{\min}\left(\tilde{\mathbf{M}}\right)\left[2 - \lambda_{\min}\left(\tilde{\mathbf{M}}\right)\right], \end{aligned}$$

where the last equality follows from the trace identity $\mathrm{tr}(\tilde{\mathbf{M}}) = \sum_k \lambda_k(\tilde{\mathbf{M}}) = 2$. Rearranging and taking square roots yields $1 - |\langle \hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_j \rangle| = \lambda_{\min}(\tilde{\mathbf{M}})$. Further, the smallest eigenvalue of the product of two matrices is bounded below by the product of the smallest eigenvalues of the individual matrices. Therefore,

$$\begin{aligned}\lambda_{\min}\left(\tilde{\mathbf{M}}\right) &\geq \lambda_{\min}\left(\mathbf{D}^{-1}\right)\lambda_{\min}\left(\mathbf{M}\right)\lambda_{\min}\left(\mathbf{D}^{-1}\right)\\ &\geq \sigma_{\min}\left(\boldsymbol{\Phi}_{\mathcal{A}\cup\{i,j\}}\right)^2 / \max\{\|\boldsymbol{\psi}_i\|^2, \|\boldsymbol{\psi}_j\|^2\}.\end{aligned}$$

The last inequality is due to Lemma A.6 and $\lambda_{\min}(\mathbf{M}) = \sigma_{\min}(\boldsymbol{\Phi}_{\mathcal{A}\cup\{i,j\}})^2$. Noting that $\max\{\|\boldsymbol{\psi}_i\|^2, \|\boldsymbol{\psi}_j\|^2\} < 1$ finishes the proof. □

The next lemma is almost identical to Lemma A.8 in its assumptions. The crucial difference is that the vectors $\boldsymbol{\psi}_i$ are projected into the column space of $\boldsymbol{\Phi}_{\mathcal{A}\setminus i}$, instead of $\boldsymbol{\Phi}_{\mathcal{A}}$. Surprisingly, the same result as in Lemma A.8 holds. Lemma A.8 and A.9 will be instrumental in providing a tight bound on the tolerable perturbation magnitude for the forward and backward algorithm, respectively.

Lemma A.9. *Let $\mathcal{A}$ be a set of k indices into columns of $\boldsymbol{\Phi}$ and $\boldsymbol{\psi}_i = (\mathbf{I} - \mathbf{P}_{\mathcal{A}\setminus i})\boldsymbol{\varphi}_i$. Then for $i \neq j$,*

$$|\langle \hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_j \rangle| \leq 1 - \sigma_{\min}\left(\boldsymbol{\Phi}_{\mathcal{A}\cup\{i,j\}}\right)^2.$$

Proof. Let $\mathcal{B} = \mathcal{A}\setminus\{i,j\}$, define $\xi_i = (\mathbf{I} - \mathbf{P}_{\mathcal{B}})\boldsymbol{\varphi}_i$, and note that $\mathbf{P}_{\mathcal{B}^\perp \mathcal{B}\cup i} = \hat{\xi}_i \hat{\xi}_i^*$, so

$$\begin{aligned}\boldsymbol{\psi}_i &= (\mathbf{I} - \mathbf{P}_{\mathcal{A}\setminus i})\boldsymbol{\varphi}_i\\ &= (\mathbf{I} - [\mathbf{P}_{\mathcal{B}} + \hat{\boldsymbol{\xi}}_j \hat{\boldsymbol{\xi}}_j^*])\boldsymbol{\varphi}_i\\ &= \boldsymbol{\xi}_i - \hat{\boldsymbol{\xi}}_j \langle \hat{\boldsymbol{\xi}}_j, \boldsymbol{\varphi}_i \rangle\\ &= \boldsymbol{\xi}_i - \hat{\boldsymbol{\xi}}_j \langle \hat{\boldsymbol{\xi}}_j, \boldsymbol{\xi}_i \rangle.\end{aligned}$$

The last equality is due to the idempotence of the projection $(\mathbf{I} - \mathbf{P}_{\mathcal{B}})$. Thus,

$$\begin{aligned}\langle \boldsymbol{\psi}_i, \boldsymbol{\psi}_j \rangle &= \langle \boldsymbol{\xi}_i - \hat{\boldsymbol{\xi}}_j \langle \hat{\boldsymbol{\xi}}_j, \boldsymbol{\xi}_i \rangle, \boldsymbol{\xi}_j - \hat{\boldsymbol{\xi}}_i \langle \hat{\boldsymbol{\xi}}_i, \boldsymbol{\xi}_j \rangle \rangle\\ &= \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_j \rangle + \langle \hat{\boldsymbol{\xi}}_j, \boldsymbol{\xi}_i \rangle \langle \hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_i \rangle \langle \hat{\boldsymbol{\xi}}_i, \boldsymbol{\xi}_j \rangle\\ &\quad - \langle \boldsymbol{\xi}_i, \hat{\boldsymbol{\xi}}_i \rangle \langle \hat{\boldsymbol{\xi}}_i, \boldsymbol{\xi}_j \rangle - \langle \hat{\boldsymbol{\xi}}_j, \boldsymbol{\xi}_j \rangle \langle \hat{\boldsymbol{\xi}}_j, \boldsymbol{\xi}_i \rangle\\ &= \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_j \rangle \left(\langle \hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_i \rangle^2 - 1 \right).\end{aligned}$$

Noting that $\|\boldsymbol{\psi}_i\|^2 = \|\boldsymbol{\xi}_i\|^2 (1 - \langle \hat{\boldsymbol{\xi}}_j, \hat{\boldsymbol{\xi}}_i \rangle^2)$. Therefore,

$$|\langle \hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_j \rangle| = |\langle \hat{\boldsymbol{\xi}}_i, \hat{\boldsymbol{\xi}}_j \rangle|.$$

Applying Lemma A.8 on the right side, and noting that $\mathcal{B}\cup\{i,j\} \subset \mathcal{A}\cup\{i,j\}$ implies $\sigma_{\min}\left(\boldsymbol{\Phi}_{\mathcal{A}\cup\{i,j\}}\right) \leq \sigma_{\min}\left(\boldsymbol{\Phi}_{\mathcal{B}\cup\{i,j\}}\right)$ finishes the proof. □

A.3. Preliminary Results for Forward Regression

We start this section by connecting separate existing results for OMP and FR in the noiseless and noisy regime. Critical to the analysis of both OMP and FR are the following quantities. Letting $\boldsymbol{\psi}_i = (\mathbf{I} - \mathbf{P}_{\mathcal{A}\setminus i})\boldsymbol{\varphi}_i$, we define

$$\rho_{\text{FR}}(\mathcal{A}, \mathcal{S}) \overset{\text{def}}{=} \frac{\max_{j\notin\mathcal{S}} |\langle \hat{\boldsymbol{\psi}}_j, \mathbf{y} \rangle|}{\max_{i\in\mathcal{S}} |\langle \hat{\boldsymbol{\psi}}_i, \mathbf{y} \rangle|} \qquad \rho_{\text{OMP}}(\mathcal{A}, \mathcal{S}) \overset{\text{def}}{=} \frac{\max_{j\notin\mathcal{S}} |\langle \boldsymbol{\psi}_j, \mathbf{y} \rangle|}{\max_{i\in\mathcal{S}} |\langle \boldsymbol{\psi}_i, \mathbf{y} \rangle|}. \tag{A.2}$$

Lemma A.10. *Suppose the support set $\mathcal{S}$ is of size k, ρ be either ρ_{FR} or ρ_{OMP} defined above, and $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{S}$. If $\rho(\emptyset, \mathcal{S}) < 1$, then ρ is a monotonically decreasing set function of its first argument. That is,*

$$\rho(\mathcal{B}, \mathcal{S}) \leq \rho(\mathcal{A}, \mathcal{S}).$$

Proof. The proof is due to (Soussen et al., 2013), Lemma 2. □

This result allows us to jointly analyze OMP and FR, since both succeed in recovering the support of a sparse signal if $\rho(\emptyset, \mathcal{S}) < 1$. The following result is a universal upper bound as a function of the Babel function values $\mu_1(k)$ and $\mu_1(k-1)$ of the matrix (see the definition of μ_1 in the main text).

Lemma A.11. *Let ρ stand for either ρ_{OMP} or ρ_{FR}, let μ_1 be the Babel function of a matrix $\mathbf{\Phi}$ with l_2-normalized columns, $\mathcal{A} \subset \mathcal{S}$ and $k = |\mathcal{S}|$. Then*

$$\rho(\emptyset, \mathcal{S}) \leq \frac{\mu_1(k)}{1 - \mu_1(k-1)}.$$

Proof. See the proof of Theorem 3.5 in (Tropp, 2004). □

Lemma A.12. *Let $\{\boldsymbol{\varphi}_i\}$ be columns of a matrix $\mathbf{\Phi}$, and μ their coherence. Then for any vector $\boldsymbol{\epsilon}$, and two indices i, j,*

$$|\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon} \rangle| + |\langle \boldsymbol{\varphi}_j, \boldsymbol{\epsilon} \rangle| \leq \sqrt{2(1+\mu)}\, \|\boldsymbol{\epsilon}\|_2.$$

Proof. First, note that $|\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon} \rangle| + |\langle \boldsymbol{\varphi}_j, \boldsymbol{\epsilon} \rangle| = \max\left\{|\langle \boldsymbol{\varphi}_i \pm \boldsymbol{\varphi}_j, \boldsymbol{\epsilon} \rangle|\right\}$. By the Cauchy-Schwartz inequality, $|\langle \boldsymbol{\varphi}_i \pm \boldsymbol{\varphi}_j, \boldsymbol{\epsilon} \rangle| \leq \|\boldsymbol{\varphi}_i \pm \boldsymbol{\varphi}_j\| \|\boldsymbol{\epsilon}\|$. The result follows by bounding $\|\boldsymbol{\varphi}_i \pm \boldsymbol{\varphi}_j\|$. To this end, note that $\|\boldsymbol{\varphi}_i \pm \boldsymbol{\varphi}_j\|^2 \leq 2\left(1 + |\langle \boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j \rangle|\right) \leq 2(1+\mu)$, where the last inequality is due to the definition of the coherence μ. □

Lemma A.13. *Suppose $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$ where $\mathbf{x}$ is k-sparse with support $\mathcal{S}$. Further, let $\mathcal{A} \subset \mathcal{S}$ and $\mathbf{r}_{\mathcal{A}} = (\mathbf{I} - \mathbf{P}_{\mathcal{A}})\mathbf{y}$. Then*

$$\max_{i \in \mathcal{S}} |\langle \boldsymbol{\varphi}_i, \mathbf{r}_{\mathcal{A}} \rangle| \geq \sigma_{\min}(\mathbf{\Phi}_{\mathcal{S}})^2 \min_{i \in \mathcal{S}} |x_i|.$$

Proof. Suppose $|\mathcal{A}| = t$ and $\mathcal{A} \subset \mathcal{S}$.

$$\begin{aligned}
\max_{i \in \mathcal{S}} |\langle \boldsymbol{\varphi}_i, \mathbf{r}_{\mathcal{A}} \rangle| &= \|\mathbf{\Phi}_{\mathcal{S}}^* (\mathbf{I} - \mathbf{P}_{\mathcal{A}})\mathbf{y}\|_\infty \\
&= \|\mathbf{\Phi}_{\mathcal{S}\setminus\mathcal{A}}^* (\mathbf{I} - \mathbf{P}_{\mathcal{A}}) \mathbf{\Phi}_{\mathcal{S}\setminus\mathcal{A}} \mathbf{x}_{\mathcal{S}\setminus\mathcal{A}}\|_\infty \\
&\geq \frac{1}{\sqrt{|\mathcal{S}\setminus\mathcal{A}|}} \|\mathbf{\Phi}_{\mathcal{S}\setminus\mathcal{A}}^* (\mathbf{I} - \mathbf{P}_{\mathcal{A}}) \mathbf{\Phi}_{\mathcal{S}\setminus\mathcal{A}} \mathbf{x}_{\mathcal{S}\setminus\mathcal{A}}\|_2 \\
&\geq \frac{\sigma_{\min}(\mathbf{\Phi}_{\mathcal{S}})^2}{\sqrt{k-t}} \|\mathbf{x}_{\mathcal{S}\setminus\mathcal{A}}\|_2 \\
&\geq \sigma_{\min}(\mathbf{\Phi}_{\mathcal{S}})^2 \min_{i \in \mathcal{S}} |x_i|.
\end{aligned}$$

The second equality comes from the fact that $(\mathbf{I} - \mathbf{P}_{\mathcal{A}})\boldsymbol{\varphi}_i = \mathbf{0}$ for all i in $\mathcal{A}$. The last inequality is due to $\|\mathbf{x}_{\mathcal{S}\setminus\mathcal{A}}\|_2^2 = \sum_{i \in \mathcal{S}\setminus\mathcal{A}} |x_i|^2 \geq (k-t) \min_{i \in \mathcal{S}} |x_i|^2$. □

Lemma A.14. *Orthogonal Matching Pursuit recovers the support set $\mathcal{S}$ of a k-sparse vector $\mathbf{x}$ provided $\rho_{OMP}(\mathcal{A}, \mathcal{S})$ in Lemma A.10 and the perturbation $\boldsymbol{\epsilon}$ of the target $\mathbf{y}$ satisfy*

$$\left[1 - \rho_{OMP}(\mathcal{A}, \mathcal{S})\right] \sigma_{\min}(\mathbf{\Phi}_{\mathcal{S}})^2 \min_{i \in \mathcal{S}} |x_i| > \max_{i \in \mathcal{S}} |\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon} \rangle| + \max_{j \notin \mathcal{S}} |\langle \boldsymbol{\varphi}_j, \boldsymbol{\epsilon} \rangle|.$$

Proof. Let $\tilde{\mathbf{y}} = \mathbf{y} + \boldsymbol{\epsilon}$, and $\mathcal{A}$ be the current active set. In order for OMP to select a column from the support $\mathcal{S}$ in the next iteration, we need to have $\max_{i \in \mathcal{S}} |\langle \boldsymbol{\varphi}_i, \tilde{\mathbf{y}} \rangle| > \max_{j \notin \mathcal{S}} |\langle \boldsymbol{\varphi}_j, \tilde{\mathbf{y}} \rangle|$. Noting that $|\langle \boldsymbol{\varphi}_i, \mathbf{y} \rangle| - |\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon} \rangle| \leq |\langle \boldsymbol{\varphi}_i, \tilde{\mathbf{y}} \rangle| \leq |\langle \boldsymbol{\varphi}_i, \mathbf{y} \rangle| + |\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon} \rangle|$, we get the sufficient condition

$$\max_{i \in \mathcal{S}} |\langle \boldsymbol{\varphi}_i, \mathbf{y} \rangle| - \max_{j \notin \mathcal{S}} |\langle \boldsymbol{\varphi}_j, \mathbf{y} \rangle| > \max_{i \in \mathcal{S}} |\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon} \rangle| + \max_{j \notin \mathcal{S}} |\langle \boldsymbol{\varphi}_j, \boldsymbol{\epsilon} \rangle|.$$

We now focus on lower bounding the left side from below. By transitivity, ensuring this lower bound is larger than the right side of the above equation is a sufficient condition for the success of the algorithm in the next iteration. By definition of $\rho(\mathcal{A}, \mathcal{S})$ in equation (A.2),

$$\max_{i \in \mathcal{S}} |\langle \boldsymbol{\varphi}_i, \mathbf{y} \rangle| - \max_{j \notin \mathcal{S}} |\langle \boldsymbol{\varphi}_j, \mathbf{y} \rangle| = \left[1 - \rho(\mathcal{A}, \mathcal{S})\right] \max_{i \in \mathcal{S}} |\langle \boldsymbol{\varphi}_i, \mathbf{y} \rangle|.$$

Applying the inequality in Lemma A.13 to the previous equation finishes the proof. □

Lemma A.15. *Forward Selection recovers the support set $\mathcal{S}$ of a k-sparse vector $\mathbf{x}$ provided $\rho_{FR}(\mathcal{A},\mathcal{S})$ in equation (A.2) and the perturbation $\boldsymbol{\epsilon}$ of $\mathbf{y}$ satisfy*

$$[1-\rho_{FR}(\mathcal{A},\mathcal{S})]\,\sigma_{\min}(\boldsymbol{\Phi}_{\mathcal{S}})^2 \min_{i\in\mathcal{S}}|x_i| > \max_{i\in\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_i,\boldsymbol{\epsilon}\rangle| + \max_{j\notin\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle|.$$

Proof. Let $\tilde{\mathbf{y}}=\mathbf{y}+\boldsymbol{\epsilon}$, and $\mathcal{A}$ be the current active set. Using Lemma B.5, we can apply similar reasoning as for OMP in Theorem A.14 to get a sufficient condition for FR to choose one of the columns in $\mathcal{S}$ in the next iteration, namely

$$[1-\rho(\mathcal{A},\mathcal{S})]\max_{i\in\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_i,\mathbf{y}\rangle| > \max_{i\in\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_i,\boldsymbol{\epsilon}\rangle| + \max_{j\notin\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle|.$$

where $\rho(\mathcal{A},\mathcal{S})$ is as in equation (A.2). We now bound the left side from below by observing that $|\langle\hat{\boldsymbol{\psi}}_i,\mathbf{y}\rangle| = |\langle\boldsymbol{\varphi}_i,\mathbf{r}_{\mathcal{A}}\rangle|/\|\boldsymbol{\psi}_i\| \geq |\langle\boldsymbol{\varphi}_i,\mathbf{r}_{\mathcal{A}}\rangle|$, where $\mathbf{r}_{\mathcal{A}}=(\mathbf{I}-\mathbf{P}_{\mathcal{A}})\mathbf{y}$. Applying the inequality of Lemma A.13 lower bounds the left side of the sufficient condition and finishes the proof. □

Lemma A.16. *Let $\mathcal{A}$ be a set of k indices into $\boldsymbol{\Phi}$ and $\boldsymbol{\psi}_i$ be either $\boldsymbol{\psi}_i=(\mathbf{I}-\mathbf{P}_{\mathcal{A}})\boldsymbol{\varphi}_i$ or $\boldsymbol{\psi}_i=(\mathbf{I}-\mathbf{P}_{\mathcal{A}\setminus i})\boldsymbol{\varphi}_i$. Then for any vector $\boldsymbol{\epsilon}$, and any two column indices i,j,*

$$|\langle\hat{\boldsymbol{\psi}}_i,\boldsymbol{\epsilon}\rangle| + |\langle\hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle| \leq \sqrt{2\left[2-\sigma_{\min}(\boldsymbol{\Phi}_{\mathcal{A}\cup\{i,j\}})^2\right]}\,\|\boldsymbol{\epsilon}\|_2.$$

Proof. This proof is similar to the one of Lemma A.12. We repeat some of the same reasoning to make the proof self contained. First, note that $|\langle\hat{\boldsymbol{\psi}}_i,\boldsymbol{\epsilon}\rangle| + |\langle\hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle| = \max\left\{|\langle\hat{\boldsymbol{\psi}}_i \pm \hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle|\right\}$. By the Cauchy-Schwartz inequality, $|\langle\hat{\boldsymbol{\psi}}_i \pm \hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle| \leq \|\hat{\boldsymbol{\psi}}_i \pm \hat{\boldsymbol{\psi}}_j\|\|\boldsymbol{\epsilon}\|$. The result follows by bounding $\|\hat{\boldsymbol{\psi}}_i \pm \hat{\boldsymbol{\psi}}_j\|$. To this end, note that $\|\hat{\boldsymbol{\psi}}_i \pm \hat{\boldsymbol{\psi}}_j\|^2 \leq 2\left(1+|\langle\hat{\boldsymbol{\psi}}_i,\hat{\boldsymbol{\psi}}_j\rangle|\right) \leq 2(2-\sigma_{\min}(\boldsymbol{\Phi}_{\mathcal{A}})^2)$, where the last inequality is due to Lemma A.8 and A.9. □

Note that Lemma A.16 improve on the trivial upper bound $2\|\boldsymbol{\epsilon}\|_2$ as long as $\boldsymbol{\Phi}_{\mathcal{A}\cup\{i,j\}}$ has linearly independent columns.

A.4. Preliminary Results for Backward Regression

Lemma A.17. *Let $\boldsymbol{\Phi}\in\mathbb{C}^{n\times m}$ be a matrix with full column rank, and let $\mathbf{x}$ be a k-sparse vector with support set $\mathcal{S}$. Let $\mathcal{A}\supset\mathcal{S}$ be the set of active indices and $\boldsymbol{\psi}_i=\boldsymbol{\varphi}_i-\mathbf{P}_{\mathcal{A}\setminus i}\boldsymbol{\varphi}_i$. Then one iteration of the backward algorithm successfully removes a column which does not belong to the true support set, provided*

$$\min_{i\in\mathcal{S}}\left|x_i\|\boldsymbol{\psi}_i\|_2 + \langle\hat{\boldsymbol{\psi}}_i,\boldsymbol{\epsilon}\rangle\right| > \min_{j\notin\mathcal{S}}\left|\langle\hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle\right|. \tag{A.3}$$

This is a necessary and sufficient condition.

Proof. A necessary and sufficient condition for the correctness of the next iteration of BE is

$$\min_{i\in\mathcal{S}}\|\mathbf{r}_{\mathcal{A}\setminus i}\|_2 > \min_{j\notin\mathcal{S}}\|\mathbf{r}_{\mathcal{A}\setminus j}\|_2.$$

Noting that $\|\mathbf{r}_{\mathcal{A}\setminus i}\|_2^2 = \|\tilde{\mathbf{y}}\|_2^2 - \langle\tilde{\mathbf{y}},\mathbf{P}_{\mathcal{A}\setminus i}\tilde{\mathbf{y}}\rangle$, the fact that $\mathbf{P}_{\mathcal{A}\setminus i}=\mathbf{P}_{\mathcal{A}}-\hat{\boldsymbol{\psi}}_i\hat{\boldsymbol{\psi}}_i^*$, and with simple algebraic manipulations, we arrive at the equivalent condition

$$\min_{i\in\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_i,\tilde{\mathbf{y}}\rangle| > \min_{j\notin\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_j,\tilde{\mathbf{y}}\rangle|. \tag{A.4}$$

Since $\boldsymbol{\psi}_i$ is orthogonal to $\mathcal{R}(\boldsymbol{\Phi}_{\mathcal{A}\setminus i})$, $\langle\hat{\boldsymbol{\psi}}_i,\mathbf{y}\rangle = \sum_{k\in\mathcal{S}}\langle\hat{\boldsymbol{\psi}}_i,\boldsymbol{\varphi}_k\rangle x_k = x_i\|\boldsymbol{\psi}_i\|_2$. Thus,

$$\langle\hat{\boldsymbol{\psi}}_i,\tilde{\mathbf{y}}\rangle = x_i\|\boldsymbol{\psi}_i\| + \langle\hat{\boldsymbol{\psi}}_i,\boldsymbol{\epsilon}\rangle.$$

Since $\boldsymbol{\psi}_j$ is orthogonal to $\mathcal{R}(\boldsymbol{\Phi}_{\mathcal{A}\setminus j})\supseteq\mathcal{R}(\boldsymbol{\Phi}_{\mathcal{S}})$, $\langle\hat{\boldsymbol{\psi}}_j,\mathbf{y}\rangle=0$, so that $\langle\hat{\boldsymbol{\psi}}_j,\tilde{\mathbf{y}}\rangle = \langle\hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle$. Therefore, (A.4) is equivalent to

$$\min_{i\in\mathcal{S}}\left|x_i\|\boldsymbol{\psi}_i\| + \langle\hat{\boldsymbol{\psi}}_i,\boldsymbol{\epsilon}\rangle\right| > \min_{j\in\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle|.$$

Since we have not made any approximations to any quantities, the condition is equivalent to the necessary and sufficient condition at the beginning of the proof. □

A.5. Probabilistic Results

In the following, let $\operatorname{erf}(x) = \int_{-\infty}^{x} e^{-y^2/2}\mathrm{d}y$ be the error function. The first result is due to a standard calculation. We report it here to compare it with our improved bounds.

Lemma A.18. *Let $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and $\boldsymbol{\Phi}$ be an $(n \times m)$-matrix with l_2-normalized columns. Then*

$$P(\|\boldsymbol{\Phi}^*\boldsymbol{\epsilon}\|_\infty < \delta) \geq 1 - m[1 - \operatorname{erf}(\delta/\sqrt{2})] \geq 1 - m\sqrt{\frac{2}{\pi}}\frac{1}{\delta}e^{-\delta^2/2}.$$

Proof. See also (Cai and Wang, 2011). As $\boldsymbol{\varphi}_i$ is l_2-normalized, $\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon}\rangle \sim \mathcal{N}(0,1)$.

$$\begin{aligned} P(\|\boldsymbol{\Phi}^*\boldsymbol{\epsilon}\|_\infty < \delta) &= 1 - P\left(\bigcup_{i=1}^{m} \{|\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon}\rangle| \geq \delta\}\right) \\ &\geq 1 - \sum P\left(|\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon}\rangle| \geq \delta\right) \\ &= 1 - 2\sum P\left(\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon}\rangle \leq -\delta\right) \\ &= 1 - m[1 - \operatorname{erf}(\delta/\sqrt{2})], \end{aligned} \tag{A.5}$$

where the last equality is due to the cumulative distribution function of the standard normal distribution being $[1 + \operatorname{erf}(x/\sqrt{2})]/2$, and $\operatorname{erf}(-x) = -\operatorname{erf}(x)$. Applying the bound $P(x > \delta) \leq e^{-\delta^2/2}/\sqrt{2\pi}\delta$ for a normal random variable $x \sim \mathcal{N}(0,1)$, to the last term in (A.5) and basic algebra finish the proof. □

Remark A.19. *Note that (Cai et al., 2009) and (Cai and Wang, 2011) provide a similar probabilistic bound for $\|\boldsymbol{\Psi}^*\boldsymbol{\epsilon}\|_\infty = \max_i |\langle \boldsymbol{\psi}_i, \boldsymbol{\epsilon}\rangle|$. After checking the proof of Lemma 5.1 in (Cai et al., 2009), we noticed that the provided bound is missing a factor of two and therefore instead bounds $\max_i \langle \boldsymbol{\psi}_i, \boldsymbol{\epsilon}\rangle$. The reason for the appearance of the factor is the penultimate equality in (A.5). As a consequence, some of the probabilistic results in (Cai and Wang, 2011) for OMP need to be adjusted minorly for correctness.*

The previous bound on the maximal absolute inner product of $\boldsymbol{\Phi}$ with $\boldsymbol{\epsilon}$ uses the subaddativity of the probability measure. The following lemmas take into account more structure of the matrix $\boldsymbol{\Phi}$ to improve on the previous result.

Lemma A.20. *Let $\boldsymbol{\Phi}$ be an $(n \times k)$ matrix with l_2-normalized columns with finite condition number $\kappa(\boldsymbol{\Phi}) < \infty$. Given $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, it holds that*

$$P(\|\boldsymbol{\Phi}^*\boldsymbol{\epsilon}\|_\infty < \delta) \geq 1 - \kappa(\boldsymbol{\Phi})^k \left[1 - \operatorname{erf}(\delta/\sqrt{2}\sigma_{\max}(\boldsymbol{\Phi}))\right]^k.$$

Proof. Note that $\boldsymbol{\Phi}^*\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{\Phi}^*\boldsymbol{\Phi}$. We will now bound the density of the distribution of $\boldsymbol{\Phi}^*\boldsymbol{\epsilon}$. To this end, note that

$$\exp\left(-\frac{\mathbf{x}^*\boldsymbol{\Sigma}^{-1}\mathbf{x}}{2}\right) \leq \exp\left(-\frac{\|\mathbf{x}\|^2}{2\lambda_{\max}(\boldsymbol{\Sigma})}\right).$$

Further, the normalization constant of the Gaussian density can be bounded by

$$\frac{1}{\sqrt{(2\pi)^k |\det \boldsymbol{\Sigma}|}} \leq \frac{1}{[2\pi\lambda_{\min}(\boldsymbol{\Sigma})]^{k/2}}.$$

Therefore, the density can be bounded above by one with a diagonal covariance,

$$\begin{aligned} P(\mathbf{x}|\boldsymbol{\Sigma}) &\overset{\text{def}}{=} (2\pi)^{-k/2} |\det \boldsymbol{\Sigma}|^{-1/2} \exp\left(-\mathbf{x}^*\boldsymbol{\Sigma}^{-1}\mathbf{x}/2\right) \\ &\leq \left(\frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})}\right)^{k/2} P(\mathbf{x}|\lambda_{\max}(\boldsymbol{\Sigma})\mathbf{I}_n) \\ &= \kappa(\boldsymbol{\Sigma})^{k/2} \prod_i P(x_i|\lambda_{\max}(\boldsymbol{\Sigma})), \end{aligned}$$

where $\kappa(\mathbf{\Sigma})$ is the condition number of $\mathbf{\Sigma}$. Assuming $\mathbf{\Phi}$ has full column rank,

$$\begin{aligned}
P(\|\mathbf{\Phi}^*\boldsymbol{\epsilon}\|_\infty < \delta) &= 1 - P\left(\bigcup_{i=1}^{k} \{|\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon}\rangle| \geq \delta\}\right) \\
&\geq 1 - \kappa(\mathbf{\Sigma})^{k/2} \prod_i P(\sigma_{\max}(\mathbf{\Phi}) \cdot |\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon}\rangle| > \delta) \\
&= 1 - \kappa(\mathbf{\Phi})^{k} \prod_i 2P(\langle \boldsymbol{\varphi}_i, \boldsymbol{\epsilon}\rangle < -\delta/\sigma_{\max}(\mathbf{\Phi})) \\
&= 1 - \kappa(\mathbf{\Phi})^{k} \left[1 - \operatorname{erf}(\delta/\sqrt{2}\sigma_{\max}(\mathbf{\Phi}))\right]^k .
\end{aligned}$$

The penultimate equality is due to $\kappa(\mathbf{\Sigma}) = \kappa(\mathbf{\Phi})^2$ and the symmetry of the standard normal distribution around zero. □

Since for sparse recovery it is commonly assumed that submatrices of $\mathbf{\Phi}$ are approximately isometric, we can use the last result to attain a bound for restrictedly isometric $\mathbf{\Phi}$ which can be much stronger than the one in Lemma A.18. This is the idea behind the next result.

Lemma A.21. *Let $\mathbf{\Phi}$ be an $(n \times m)$-matrix with l_2-normalized columns with Babel function μ_1 and $d = \lceil m/k \rceil$. Given $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$,*

$$P(\|\mathbf{\Phi}^*\boldsymbol{\epsilon}\|_\infty < \delta) \geq 1 - d\sqrt{\frac{1+\mu_1(k)}{1-\mu_1(k)}}^{k} \left[1 - \operatorname{erf}(\delta/\sqrt{2[1+\mu_1(k)]})\right]^k .$$

Proof. We first divide the dictionary into $d = \lceil m/k \rceil$ sets of k columns each, thereby accounting for $(m \bmod k)$ columns twice. Let the indices of the ith group be $\mathcal{S}_i = \{z \bmod m | (i-1)k < z \leq ik\}$ for $1 \leq i \leq d$.

$$\begin{aligned}
P(\|\mathbf{\Phi}^*\boldsymbol{\epsilon}\|_\infty < \delta) &= 1 - P\left(\bigcup_{i=1}^{k} \left\{\|\mathbf{\Phi}_{\mathcal{S}_i}^*\boldsymbol{\epsilon}\|_\infty \geq \delta\right\}\right) \\
&\geq 1 - \sum P\left(\|\mathbf{\Phi}_{\mathcal{S}_i}^*\boldsymbol{\epsilon}\|_\infty \geq \delta\right) \\
&\geq 1 - d\kappa(\mathbf{\Phi}_{\mathcal{S}_i})^k \left[1 - \operatorname{erf}(\delta/\sigma_{\max}(\mathbf{\Phi}_{\mathcal{S}_i}))\right]^k .
\end{aligned}$$

The last inequality is an application of the previous lemma. Lemma A.5 implies $\sigma_{\max}(\mathbf{\Phi})^2 \leq 1 + \mu_1(k)$ which means $\kappa(\mathbf{\Phi})^2 \leq (1+\mu_1(k))/(1-\mu_1(k))$. Plugging the bounds into Lemma A.20 finishes the proof. □

B. Results for Relevance Matching Pursuit

Recall from the main text that

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) &\stackrel{\text{def}}{=} \log \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}|\boldsymbol{\gamma})d\mathbf{x} \\ &= -\mathbf{y}^*\mathbf{C}^{-1}\mathbf{y} + \log|\mathbf{C}| - n\log(2\pi)\end{aligned} \tag{B.1}$$

is the logarithm of the marginal likelihood. Further, $\mathbf{C} = (\sigma^2\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^*)$ is the covariance of the marginal distribution, $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$ is the prior covariance of the weights $\mathbf{x}$, $\mathcal{A} = \{i|\gamma_i \neq 0\}$ is the active set, $s_i \stackrel{\text{def}}{=} \boldsymbol{\varphi}_i\mathbf{C}_{\mathcal{A}\setminus i}^{-1}\mathbf{y}$ and $q_i \stackrel{\text{def}}{=} \boldsymbol{\varphi}_i\mathbf{C}_{\mathcal{A}\setminus i}^{-1}\boldsymbol{\varphi}_i$.

B.1. Properties of the marginal likelihood

Given q_i, s_i and keeping all other parameters constant, the optimal prior variance of the ith feature γ_i is given by

$$\begin{aligned}\gamma_i &= \frac{q_i^2 - s_i}{s_i^2} \quad &\text{if } q_i^2 > s_i, \\ \gamma_i &= 0 \quad &\text{if } q_i^2 \leq s_i.\end{aligned} \tag{B.2}$$

According to (Tipping and Faul, 2003), the increase in the logarithm of the marginal likelihood upon adding an inactive $\boldsymbol{\varphi}_i$ and setting its prior variance γ_i to its optimal value via (B.2) is

$$\Delta_{\text{add}}(i) \stackrel{\text{def}}{=} \left(\frac{q_i^2 - s_i}{s_i} + \log\frac{s_i}{q_i^2}\right)/2. \tag{B.3}$$

Similarly, setting a γ_i of an active $\boldsymbol{\varphi}_i$ to its optimal value changes the marginal likelihood by

$$\Delta_{\text{update}}(i) \stackrel{\text{def}}{=} \left(\frac{Q_i^2 - S_i}{S_i + (\tilde{\gamma}_i - \gamma_i)^{-1}} - \log\left[1 + S_i\left(\tilde{\gamma}_i - \gamma_i\right)^{-1}\right]\right)/2, \tag{B.4}$$

where $S_i \stackrel{\text{def}}{=} \boldsymbol{\varphi}_i\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{y}$ and $Q_i \stackrel{\text{def}}{=} \boldsymbol{\varphi}_i\mathbf{C}_{\mathcal{A}}^{-1}\boldsymbol{\varphi}_i$, and γ_i and $\tilde{\gamma}_i$ are the current and optimal prior variances in (B.4), respectively. Δ_{update} is used in the update step of RMP to pick the active feature with the maximal increase in the marginal likelihood.

B.2. Statements and proofs

Lemma B.1. *The optimum of the marginal likelihood with respect to γ_i occurs at a non-zero value if and only if*

$$|\langle\tilde{\boldsymbol{\varphi}}_i|\mathbf{r}_{\mathcal{A}\setminus i,\sigma}\rangle| > \sigma,$$

where $\tilde{\boldsymbol{\varphi}}_i \stackrel{\text{def}}{=} \boldsymbol{\varphi}_i/\|\boldsymbol{\varphi}_i\|_{\mathbf{R}_{\mathcal{A}\setminus i,\sigma}}$, $\|\boldsymbol{\varphi}_i\|_{\mathbf{R}_{\mathcal{A}\setminus i,\sigma}}$ is the energetic norm $(\boldsymbol{\varphi}_i^*\mathbf{R}_{\mathcal{A}\setminus i,\sigma}\boldsymbol{\varphi}_i)$ of $\boldsymbol{\varphi}_i$, and $\mathcal{A}$ is the active set.

Proof. First, note that via the Woodbury identity, we get

$$\mathbf{R}_{\mathcal{A},\sigma} \stackrel{\text{def}}{=} \sigma^2\mathbf{C} = \mathbf{I} - \boldsymbol{\Phi}_{\mathcal{A}}(\sigma^2{\boldsymbol{\Gamma}_{\mathcal{A}}}^{-1} + \boldsymbol{\Phi}_{\mathcal{A}}^T\boldsymbol{\Phi}_{\mathcal{A}})^{-1}\boldsymbol{\Phi}_{\mathcal{A}}^T = \mathbf{I} - \boldsymbol{\Phi}_{\mathcal{A}}[\sigma^{-2}\boldsymbol{\Sigma}_{\mathcal{A}}]\boldsymbol{\Phi}_{\mathcal{A}}^T. \tag{B.5}$$

Further, if the optimum of the likelihood with respect to γ_i occurs at a non-zero value, we must have $1 < q_i^2/s_i$. Thus,

$$1 < \frac{q_i^2}{s_i} = \frac{\langle\boldsymbol{\varphi}_i|\mathbf{C}_{\mathcal{A}\setminus i}\mathbf{y}\rangle^2}{\langle\boldsymbol{\varphi}_i|\mathbf{C}_{\mathcal{A}\setminus i}\boldsymbol{\varphi}_i\rangle} = \frac{\langle\boldsymbol{\varphi}_i|\mathbf{R}_{\mathcal{A}\setminus i,\sigma}\mathbf{y}\rangle^2}{\sigma^2\langle\boldsymbol{\varphi}_i|\mathbf{R}_{\mathcal{A}\setminus i,\sigma}\boldsymbol{\varphi}_i\rangle} = \frac{\langle\boldsymbol{\varphi}_i|\mathbf{R}_{\mathcal{A}\setminus i,\sigma}\mathbf{y}\rangle^2}{\sigma^2\|\boldsymbol{\varphi}_i\|^2_{\mathbf{R}_{\mathcal{A}\setminus i,\sigma}}} = \frac{1}{\sigma^2}\langle\tilde{\boldsymbol{\varphi}}_i|\mathbf{r}_{\mathcal{A}\setminus i,\sigma}\rangle^2, \tag{B.6}$$

where $\tilde{\boldsymbol{\varphi}}_i = \boldsymbol{\varphi}_i/\|\boldsymbol{\varphi}_i\|_{\mathbf{R}_{\mathcal{A}\setminus i,\sigma}}$, and $\mathbf{r}_{\mathcal{A},\sigma} = \mathbf{y} - \boldsymbol{\Phi}_{\mathcal{A}}\boldsymbol{\mu}_{\mathcal{A},\sigma}$ is the residual associated with the posterior mean of the coefficients $\boldsymbol{\mu}_{\mathcal{A},\sigma}$ under the SBL model. Taking square-roots and multiplying through by σ yields the result. □

Lemma B.2. *Let $\Delta_{add}(i)$ be the change in the marginal likelihood upon setting an inactive feature's prior variance γ_i to its optimal value via equation (B.2). Then*

$$\arg\max_{i\notin\mathcal{A}} \Delta_{add}(i) = \arg\max_{i\notin\mathcal{A}} |\langle\tilde{\boldsymbol{\varphi}}_i|\mathbf{r}_{\mathcal{A},\sigma}\rangle|.$$

where $\tilde{\boldsymbol{\varphi}}_i = \boldsymbol{\varphi}_i/\|\boldsymbol{\varphi}_i\|_{\mathbf{R}_{\mathcal{A},\sigma}}$, and $\mathcal{A}$ is the active set.

Proof. The increase in the log marginal likelihood by adding an feature, and setting its prior variance to the optimal value is given by equation (B.3). By rearranging, we can write

$$2\Delta_{\text{add}}(i) = \frac{q_i^2 - s_i}{s_i} + \log\frac{s_i}{q_i^2} = \left(\frac{q_i^2}{s_i}\right) - \log\left(\frac{q_i^2}{s_i}\right) - 1. \tag{B.7}$$

Evidently, the increase $\Delta_{\text{add}}(i)$ is a univariate function of the fraction q_i^2/s_i only. The function is $f(x) = x - \log x - 1$, which is strictly increasing for $x > 1$. Therefore, choosing the feature with the largest marginal likelihood increase corresponds to choosing the feature with the largest q_i^2/s_i value above 1. If no ratio is above 1, no feature is added. According to Lemma B.1, $q_i^2/s_i = \langle\tilde{\boldsymbol{\varphi}}_i|\mathbf{r}_{\mathcal{A}\backslash i,\sigma}\rangle/\sigma^2$. Noting that $i \notin \mathcal{A}$, implies $\mathcal{A}\backslash i = \mathcal{A}$ finishes the proof. □

Lemma B.3. *As $\delta_{\mathcal{L}} \to 0$, the $\boldsymbol{\gamma}$ returned by RMP$_\sigma$ constitutes a local maximum of the marginal likelihood.*

Proof. This is essentially due to Faul and Tipping (2002). To elaborate, note that RMP can only terminate if no update yields an improvement above the threshold $\delta_{\mathcal{L}}$, nor can a feature be added or deleted without decreasing the marginal likelihood. Indeed, after the second inner loop breaks, we are guaranteed to have $\Delta_{\text{update}}(i) < \delta_{\mathcal{L}}$ for all $i \in \mathcal{A}$. Further, the coordinate ascent updates to γ_i do provably converge to a joint maximum, not merely a stationary point (Faul and Tipping, 2002). These facts imply that as we move $\delta_{\mathcal{L}} \to 0^+$, the results of the algorithms converge to a local maximum of the marginal likelihood at which $\Delta_{\text{update}}(i) = 0$ for all $i \in \mathcal{A}$ and no features can be added or deleted without decreasing the likelihood. □

Lemma B.4. *Assume the columns of $\boldsymbol{\Phi}_{\mathcal{A}}$, are linearly independent. Then*

$$\mathbf{R}_{\mathcal{A}} \overset{def}{=} (\mathbf{I} - \boldsymbol{\Phi}_{\mathcal{A}}\boldsymbol{\Phi}_{\mathcal{A}}^{+}) = \lim_{\sigma\to 0^+} \mathbf{R}_{\mathcal{A},\sigma}.$$

Proof. Using a standard limit relation for the pseudoinverse (Golub and Van Loan, 2012),

$$\lim_{\sigma\to 0^+} \mathbf{R}_{\mathcal{A},\sigma} = \mathbf{I} - (\boldsymbol{\Phi}_{\mathcal{A}}\boldsymbol{\Gamma}_{\mathcal{A}}^{\frac{1}{2}})(\boldsymbol{\Phi}_{\mathcal{A}}\boldsymbol{\Gamma}_{\mathcal{A}}^{\frac{1}{2}})^{+}. \tag{B.8}$$

If further the columns $\boldsymbol{\Phi}_{\mathcal{A}}$ of $\boldsymbol{\Phi}$ for which the γ_i are non-zero are linearly independent, $(\boldsymbol{\Phi}_{\mathcal{A}}\boldsymbol{\Gamma}_{\mathcal{A}}^{1/2})^{+} = \boldsymbol{\Gamma}_{\mathcal{A}}^{-1/2}\boldsymbol{\Phi}_{\mathcal{A}}^{+}$, so that $\lim_{\sigma\to 0^+} \mathbf{R}_{\mathcal{A},\sigma} = \mathbf{I} - \boldsymbol{\Phi}_{\mathcal{A}}\boldsymbol{\Phi}_{\mathcal{A}}^{+}$. □

Lemma B.5. *Let $\mathbf{r}_{\mathcal{A}}$ be the least-squares residual associated with a feature set $\mathcal{A}$. Then*

$$\begin{aligned} \|\mathbf{r}_{\mathcal{A}}\|_2^2 - \|\mathbf{r}_{\mathcal{A}\cup i}\|_2^2 &= |\langle\boldsymbol{\varphi}_i, \mathbf{r}_{\mathcal{A}}\rangle|^2/\|\boldsymbol{\varphi}_i\|_{\mathbf{R}_{\mathcal{A}}}^2, \text{and} \\ \|\mathbf{r}_{\mathcal{A}\backslash i}\|_2^2 - \|\mathbf{r}_{\mathcal{A}}\|_2^2 &= |\langle\boldsymbol{\varphi}_i, \mathbf{r}_{\mathcal{A}}\rangle|^2/\|\boldsymbol{\varphi}_i\|_{\mathbf{R}_{\mathcal{A}\backslash i}}^2. \end{aligned} \tag{B.9}$$

Proof. We prove the first equality and note that the second equality can be proven similarly. Let $\boldsymbol{\psi}_i = \mathbf{R}_{\mathcal{A}}\boldsymbol{\varphi}_i$. Then the projection into the orthogonal complement $\text{col}(\boldsymbol{\Phi}_{\mathcal{A}})$ in $\text{col}(\boldsymbol{\Phi}_{\mathcal{A}\cup i})$ is equal to $\hat{\boldsymbol{\psi}}_i\hat{\boldsymbol{\psi}}_i^*$. Therefore,

$$\begin{aligned} \|\mathbf{r}_{\mathcal{A}\cup i}\|^2 &= \|(\mathbf{I} - \mathbf{P}_{\mathcal{A}\cup i})\mathbf{y}\|^2 \\ &= \|(\mathbf{I} - [\mathbf{P}_{\mathcal{A}} + \hat{\boldsymbol{\psi}}_i\hat{\boldsymbol{\psi}}_i^*])\mathbf{y}\|^2 \\ &= \|\mathbf{r}_{\mathcal{A}}\|^2 - |\langle\hat{\boldsymbol{\psi}}_i, \mathbf{y}\rangle|^2 \\ &= \|\mathbf{r}_{\mathcal{A}}\|^2 - |\langle\boldsymbol{\varphi}_i, \mathbf{R}_{\mathcal{A}}\mathbf{y}\rangle|^2/\|\boldsymbol{\varphi}\|_{\mathbf{R}_{\mathcal{A}}}^2. \end{aligned}$$

□

C. Results For Stepwise Regression

C.1. Forward Regression

Theorem C.1. *Suppose the columns of $\mathbf{\Phi}$ are l_2-normalized. Orthogonal Matching Pursuit and Forward Regression recover the support $\mathcal{S}$ of a k-sparse vector in k iterations provided the Babel function μ_1 of $\mathbf{\Phi}$ and the perturbation $\boldsymbol{\epsilon}$ of the target $\mathbf{y}$ satisfy*

$$\frac{1-2\mu_1(k)}{\sqrt{2[1+\mu_1(k)]}} \min_{i\in\mathcal{S}} |x_i| \geq \|\boldsymbol{\epsilon}\|_2.$$

Proof. Consider the left side of the inequality of Lemma A.14 and Lemma A.15. According to Lemma A.11, both ρ_{OMP} and ρ_{FR} are upper bounded by $\frac{\mu_1(k)}{1-\mu_1(k-1)}$. Further, Lemma A.5 implies $1-\mu_1(k) \leq \sigma_{\min}(\mathbf{\Phi}_{\mathcal{S}})^2$, and μ_1 is increasing. Therefore,

$$\begin{aligned}(1-\rho)\sigma_{\min}(\mathbf{\Phi}_{\mathcal{S}})^2 &\geq \left[1-\frac{\mu_1(k)}{1-\mu_1(k-1)}\right](1-\mu_1(k-1)) \\ &= 1-\mu_1(k-1)-\mu_1(k) \\ &\geq 1-2\mu_1(k).\end{aligned}$$

Regarding OMP, the right side of the inequality in Lemma A.14 is bounded by $\sqrt{2(1+\mu)}\|\boldsymbol{\epsilon}\|$, according to Lemma A.12. Similarly, regarding FR, the right side of the inequality in Lemma A.15 is bounded by $\sqrt{2[1+\mu_1(k)]}\|\boldsymbol{\epsilon}\|$, according Lemma A.16. As the Babel function is increasing, the latter bound holds for both OMP and FR. This finishes the proof. □

Theorem C.2. *Suppose $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\sigma^2)$, and let $\delta = [1/2-\mu_1(k)]\min_{i\in\mathcal{S}}|x_i|/\sigma$. Orthogonal Matching Pursuit and Forward Regression recover the support of a k-sparse signal with probability exceeding*

$$1-\left\lceil\frac{m}{k}\right\rceil\left(\frac{1+\kappa_1(k)}{1-\mu_1(2k)}\right)^{k/2}\left(1-\operatorname{erf}(\delta/\sqrt{2\kappa_1(k)})\right)^k$$

where $\kappa_1(k) \overset{def}{=} (1+\mu_1(2k))/(1-\mu_1(k))$. Assuming $\mu_1(2k) < 1/2$, this holds with probability exceeding

$$1-\left\lceil\frac{m}{k}\right\rceil\left[2\sqrt{2}\left(1-\operatorname{erf}\left(\delta/\sqrt{6}\right)\right)\right]^k > 1-\left\lceil\frac{m}{k}\right\rceil\left[\frac{4}{\sqrt{\pi}\delta}e^{-\delta^2/6}\right]^k.$$

Proof. For OMP, the right side of the inequality in Lemma A.14 can be bounded by

$$\max_{i\in\mathcal{S}}|\langle\boldsymbol{\varphi}_i,\boldsymbol{\epsilon}\rangle| + \max_{j\notin\mathcal{S}}|\langle\boldsymbol{\varphi}_j,\boldsymbol{\epsilon}\rangle| \leq 2\|\mathbf{\Phi}^*\boldsymbol{\epsilon}\|_\infty.$$

Noting that $\langle\boldsymbol{\varphi}_i,\boldsymbol{\epsilon}\rangle$ has standard deviation σ, we can apply Lemma A.21 to the right hand side of the last inequality and get a lower bound on the probability with which OMP recovers the correct support:

$$1-\left\lceil\frac{m}{k}\right\rceil\left(\frac{1+\mu_1(k)}{1-\mu_1(k)}\right)^{k/2}\left[1-\operatorname{erf}(\delta/\sqrt{2[1+\mu_1(k)]})\right]^k.$$

Similarly for Forward Regression, the right side of the inequality in Lemma A.15 can be bounded by

$$\max_{i\in\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_i,\boldsymbol{\epsilon}\rangle| + \max_{j\notin\mathcal{S}}|\langle\hat{\boldsymbol{\psi}}_j,\boldsymbol{\epsilon}\rangle| \leq 2\|\hat{\mathbf{\Psi}}^*_{\mathcal{A}^c}\boldsymbol{\epsilon}\|_\infty,$$

where, for any index set $\mathcal{B}$, we define $\mathbf{\Psi}_{\mathcal{B}} = (\mathbf{I}-\mathbf{P}_{\mathcal{A}})\mathbf{\Phi}_{\mathcal{B}}$ and $\hat{\mathbf{\Psi}}_{\mathcal{A}^c}$ is the result of normalizing the columns of $\mathbf{\Psi}_{\mathcal{A}^c}$. In order to use Lemma A.21 on the right hand side of the last inequality, we first need to bound the singular values of the submatrices of $\hat{\mathbf{\Psi}}_{\mathcal{A}^c}$.

To this end, let $\mathcal{B}\subset\mathcal{A}^c$ and $|\mathcal{B}| = |\mathcal{A}| = k$. Using Lemma A.6 and A.7,

$$\sigma_{\min}(\hat{\mathbf{\Psi}}_{\mathcal{B}}) \geq \sigma_{\min}(\mathbf{\Psi}_{\mathcal{B}})/\max_{i\in\mathcal{B}}\|\boldsymbol{\psi}_i\| \geq \sigma_{\min}((\mathbf{I}-\mathbf{P}_{\mathcal{A}})\mathbf{\Phi}_{\mathcal{B}}) \geq \sigma_{\min}(\mathbf{\Phi}_{\mathcal{A}\cup\mathcal{B}}) \geq 1-\mu_1(2k).$$

In a similar vain,

$$\sigma_{\max}(\hat{\mathbf{\Psi}}_{\mathcal{B}}) \leq \sigma_{\max}(\mathbf{\Phi}_{\mathcal{A}\cup\mathcal{B}})/\sigma_{\min}(\mathbf{\Phi}_{\mathcal{B}}) \leq \kappa_1(k) \stackrel{\text{def}}{=} \frac{1+\mu_1(2k)}{1-\mu_1(k)}.$$

Using these bounds on the extremal singular values in place of $1 \pm \mu_1(k)$ in the proof of Lemma A.21, we attain the first inequality of the result (Theorem C.2). As this inequality is slightly looser than the one derived above for OMP alone, it holds for both algorithms. The second inequality of the result follows by substituting $1/2$ for $\mu_1(2k)$ into the expression and by noting that $\mu_1(k) \leq \mu_1(2k)$, which follows from the known property that for all $k \geq 1$, $\mu_1(k) \leq k\mu$. In that case, $\kappa_1(k) = (1+\mu_1(2k))/(1-\mu_1(k)) < (3/2)/(1/2) = 3$ and thus $\sqrt{(1+\kappa_1(k))/(1-\mu_1(k))} < 2\sqrt{2}$. The last inequality of the second equation is due to the standard Gaussian tail bound already employed in Lemma A.18. □

C.2. Backward Regression

Theorem C.3. *Suppose $\mathbf{\Phi}$ has full column rank. Then Backward Regression recovers the support $\mathcal{S}$ of a k-sparse m-dimensional vector $\mathbf{x}$ in $m-k$ steps if*

$$\frac{\sigma_{\min}(\mathbf{\Phi})}{\sqrt{2[2-\sigma_{\min}(\mathbf{\Phi})^2]}} \min_{i\in\mathcal{S}} |x_i| > \|\boldsymbol{\epsilon}\|_2,$$

where $\sigma_{\min}(\mathbf{\Phi})$ is the smallest singular value of $\mathbf{\Phi}$.

Proof. The left side of equation (A.3) in Lemma A.17 can be bounded below by

$$\min_{i\in\mathcal{S}} \left[|x_i|\|\boldsymbol{\psi}_i\| - \left|\langle \hat{\boldsymbol{\psi}}_i, \boldsymbol{\epsilon}\rangle\right| \right] \geq \min_{i\in\mathcal{S}} |x_i|\|\boldsymbol{\psi}_i\| - \max_{i\in\mathcal{S}} \left|\langle \hat{\boldsymbol{\psi}}_i, \boldsymbol{\epsilon}\rangle\right|.$$

A sufficient condition for the one-step success of the algorithm is then

$$\begin{aligned} \min_{i\in\mathcal{S}} |x_i|\|\boldsymbol{\psi}_i\| &\geq \sigma_{\min}(\mathbf{\Phi}_{\mathcal{A}}) \min_{i\in\mathcal{S}} |x_i| \\ &> \sqrt{2[2-\sigma_{\min}(\mathbf{\Phi}_{\mathcal{A}})]}\|\boldsymbol{\epsilon}\|_2 \geq \max_{i\in\mathcal{S}} \left|\langle \hat{\boldsymbol{\psi}}_i, \boldsymbol{\epsilon}\rangle\right| + \min_{j\notin\mathcal{S}} \left|\langle \hat{\boldsymbol{\psi}}_j, \boldsymbol{\epsilon}\rangle\right|. \end{aligned} \tag{C.1}$$

The lower bound of $\|\boldsymbol{\psi}_i\| \geq \sigma_{\min}(\mathbf{\Phi}_{\mathcal{A}})$ is a direct consequence of Lemma A.7. The last upper bound is due to Lemma A.16. The second inequality is forced to guarantee the one-step success of the algorithm. Since $\sigma_{\min}(\mathbf{\Phi}_{\mathcal{A}}) \geq \sigma_{\min}(\mathbf{\Phi})$ for all submatrices $\mathbf{\Phi}_{\mathcal{A}}$ of $\mathbf{\Phi}$, every iteration is guaranteed to successfully remove an irrelevant feature if the inequality in the statement of the theorem holds, until the correct support $\mathcal{S}$ is recovered. □

Corollary C.4. *Suppose $\mathbf{\Phi}_{\mathcal{A}}$ has full column rank, $|\mathcal{A}| = k$, and $\mathcal{S} \subset \mathcal{A}$. Then Backward Regression recovers the correct support set in $|\mathcal{A}| - |\mathcal{S}|$ iterations, provided*

$$\sqrt{\frac{1-\mu_1(k)}{2[1+\mu_1(k)]}} \min_{i\in\mathcal{S}} |x_i| > \|\boldsymbol{\epsilon}\|_2.$$

Proof. First, Lemma A.5 implies $\sigma^2_{\min}(\mathbf{\Phi}_{\mathcal{A}}) > 1-\mu_1(k-1)$ for all index sets $\mathcal{A}$ of size k. Since μ_1 is increasing, we can bound

$$\frac{\sigma_{\min}(\mathbf{\Phi})}{\sqrt{2[2-\sigma_{\min}(\mathbf{\Phi})^2]}} \geq \sqrt{\frac{1-\mu_1(k)}{2[1+\mu_1(k)]}}.$$

□

Corollary C.5. *Let $\mathbf{x}_k$ be the vector that achieves the smallest residual norm $\|\mathbf{y}-\mathbf{\Phi}\mathbf{x}\|$ among all vectors $\mathbf{x}$ with k or fewer non-zero elements. If the associated residual $\mathbf{r}_k \stackrel{\text{def}}{=} \mathbf{y}-\mathbf{\Phi}\mathbf{x}_k$ satisfies the bound in Theorem C.3 in place of $\boldsymbol{\epsilon}$, Backward Regression recovers $\mathbf{x}_k$, or equivalently, solves the subset selection problem to optimality.*

Proof. The main idea is to reduce the problem to the sparse recovery problem. That is, letting $\boldsymbol{\epsilon} \leftarrow \mathbf{r}_k$, the problem becomes indistinguishable from sparse recovery with a k-sparse coefficient vector $\mathbf{x}_k$ with support $\mathcal{S}$. If the conditions of Theorem C.3 hold, the result guarantees the recovery of the support $\mathcal{S}$ of $\mathbf{x}_k$. Since, by definition, $\mathbf{x}_k$ attains the minimal residual among all k-sparse vectors, and $\mathbf{\Phi}_{\mathcal{S}}\mathbf{\Phi}_{\mathcal{S}}^{+}(\mathbf{y}+\boldsymbol{\epsilon})$ reaches the minimal residual among all vectors with support $\mathcal{S}$, the backward algorithm returns *precisely* $\mathbf{x}_k$. □

D. Numerical Experiments

D.1. Implementation Details and Experimental Setup

We made all our implementations publicly available via the `CompressedSensing.jl` package by the publication date of ICML 2021 to function as a platform for sparsity-inducing algorithms. The implementations are contained in the src folder. OMP is in matchingpursuit.jl, FR is in forward.jl, BR is in backward.jl, RMP and FoBa are in stepwise.jl, the SBL algorithms are in sbl.jl, and BP is in basispursuit.jl

For the sake of reproducibility, we additionally attached the folder ICML2021, which contains all code to run the experiments and the results stored in H5 files.

D.2. Estimation of the noise variance

Here, we discuss a potential addition to RMP: the estimation of the noise variance σ^2, which is one of the advantages of the probabilistic framework over traditional compressed sensing and feature selection methodologies. To this end, setting the derivative of the marginal likelihood with respect to σ^2 to zero yields the update $\sigma^2 \leftarrow \|\mathbf{y} - \mathbf{\Phi}\boldsymbol{\mu}\|^2/(n + \sum_i \gamma_i)$, where n is the number of rows of $\mathbf{\Phi}$ (Tipping, 2001). This update can be applied at any point during the execution of the algorithm, while maintaining the same local convergence guarantees. However, adding this update to the loop of any coordinate-ascent algorithm for SBL requires the factorization of $\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{\Phi}_{\mathcal{A}}\mathbf{\Gamma}\mathbf{\Phi}_{\mathcal{A}}^*$ from scratch every time σ^2 is changed. In contrast, coordinate-wise updates of $\mathbf{\Gamma}$ result in efficient low-rank updates to $\mathbf{C}$. Thus, re-estimation of σ^2 in the loop would increase the computational complexity per iteration from $O(nk)$ to $O(nk^2)$ where k is the size of the active set $\mathcal{A}$. For this reason, we propose an algorithm that alternates the application of the σ-update and a run of RMP, warm-started at the existing parameters and run to convergence. By running experiments with this algorithm, we noticed that the success of the estimation depends on the sampling ratio n/m. Table 1 shows this dependence by reporting statistics of the estimated value of σ^2 after convergence of this procedure over 128 independently generated instances with $m = 128$ and $k = 4$ fixed. The observations were corrupted with Gaussian noise with $\sigma^2 = 1\text{e}{-4}$. For low sampling fractions, i.e. the domain of compressed sensing, neither mean nor median are close to the true value, indicating convergence to local minima different from the ground truth, despite having initialized the variance very close to the ground truth at $(2\sigma)^2$. For high sampling fractions, the median estimate gets increasingly close to the ground truth, though the variance in the results stays high. To stabilize the estimation, it is common to add an inverse gamma prior on the noise variance. We also ran experiments using ranges of different prior values but found that, as long as the prior is just relatively vague, the variability of the results remains high. If the noise variance really is completely unknown, one might have to resort to a fully Bayesian approach using MCMC sampling to get representative values of the noise variance. However, such an approach is firmly outside the scope of the current work. We also note that with the exception of SBL, none of the other sparsity-promoting algorithms offer this feature.

Table 1: Statistics of σ^2 estimation via marginal likelihood optimization computed over 128 independent instances with $m = 128$ and $k = 4$ fixed. The ground truth is $\sigma^2 = 1\text{e}{-4}$.

n/m	1/4	1/2	1	2	4	8
mean	5.18e−01	5.16e−01	3.96e−01	2.58e−01	2.34e−01	1.82e−01
median	3.72e−06	3.96e−06	4.19e−05	7.69e−05	9.07e−05	9.48e−05
std	1.31e+00	1.03e+00	1.19e+00	8.13e−01	8.80e−01	9.37e−01

References

Cai, T. T. and Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688.

Cai, T. T., Xu, G., and Zhang, J. (2009). On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 55(7):3388–3397.

Couvreur, C. and Bresler, Y. (2000). On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM Journal on Matrix Analysis and Applications*, 21(3):797–808.

Faul, A. C. and Tipping, M. E. (2002). Analysis of sparse bayesian learning. In *Advances in neural information processing systems*, pages 383–389.

Golub, G. and Van Loan, C. (2012). *Matrix Computations*. Matrix Computations. Johns Hopkins University Press.

Soussen, C., Gribonval, R., Idier, J., and Herzet, C. (2013). Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares. *IEEE Transactions on Information Theory*, 59(5):3158–3174.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244.

Tipping, M. E. and Faul, A. C. (2003). Fast marginal likelihood maximisation for sparse bayesian models. In *AISTATS*.

Tropp, J. A. (2004). Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242.