SPARSE RECOVERY VIA BOOTSTRAPPING: COLLABORATIVE OR INDEPENDENT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse regression problems have traditionally been solved using all available measurements simultaneously. However, this approach fails in challenging scenarios such as when the noise level is high or there are missing data / adversarial samples.

We propose JOBS (Joint-Sparse Optimization via Bootstrap Samples) – a *collaborative* sparse-regression framework on bootstrapped samples from the pool of available measurements via a joint-sparse constraint to ensure support consistency. In comparison to traditional bagging which solves sub-problems in an *independent* fashion across bootstrapped samples, JOBS achieves state-of-the-art performance with the added advantage of having a sparser solution while requiring a lower number of observation samples.

Analysis of theoretical performance limits is employed to determine critical optimal parameters: the number of bootstrap samples K and the number of elements L in each bootstrap sample. Theoretical results indicate a better bound than Bagging (i.e. higher probability of achieving the same or better performance). Simulation results are used to validate this parameter selection. JOBS is robust to adversarial samples that fool the baseline method, as shown by better generalization in an image reconstruction task where the adversary has similar occlusions or alignment as the test sample. Furthermore, JOBS also improves discriminative performance in a facial recognition task in a sparse-representation-based classification setting.

1 INTRODUCTION

In compressed sensing (CS) and sparse regression, a classic linear inverse solution via least squares plus a sparsity-promoting penalty term has been extensively studied. Sparse regression is important for feature selection, reducing over-fitting, and representation learning. and there are rich variants that solve important problems such as dictionary learning (Duarte-Carvajalino & Sapiro, 2009), matrix completion (Candès & Recht, 2009), Robust Principle Component Analysis (Candès et al., 2011), matrix factorization (Lee & Seung, 2001), and sparse neural networks (Alvarez & Salzmann, 2016).

Mathematically speaking, let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be the sensing matrix, $\boldsymbol{x} \in \mathbb{R}^n$ contains the sparse codes with very few non-zero entries, $\boldsymbol{z}$ is a noise vector with low bounded energy, and $\boldsymbol{y} \in \mathbb{R}^m$ be the measurement vector, commonly generated by a linear model with measurement noise: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}$. The ℓ_1 norm minimization is the most common strategy, also known as LASSO (Tibshirani, 1996) or Basis Pursuit denoising (Chen et al., 2001).

The performance of ℓ_1 minimization has been thoroughly studied in the CS literature (Cohen et al., 2009; Candes, 2008; Candes et al., 2006; Donoho, 2006; Candess & Romberg, 2007), including the correctness and robustness based on the Null Space Property (NSP) (Cohen et al., 2009) and the Restricted Isometry Property (RIP) (Candes, 2008; Candes et al., 2006) and mild sufficient conditions on random matrices with sufficient sample complexity to obtain bounded reconstruction error with high probability (Candes, 2008).

Even though the baseline ℓ_1 min., works pretty well in many applications, Unfortunately, its performance suffers in challenging, high noise cases. Moreover it has trouble with partially missing and/or severely corrupted samples. Additionally, it is not robust against adversarial samples (illustrated in Figure 1) which are similar to the test case but are from a different class. Adversarial samples can be caused by lack of variation in the training data, and algorithms that can overcome these samples exhibit better generalization.

Bagging, a classic method for regression and classification tasks, has shown its robustness in high noise cases (Breiman, 1996). In this paper, we use *Bagging* to refer to employing Bagging procedure in sparse recovery. To obtain the Bagging solution, the same objective function is solved multiple times independently from bootstrap (Efron, 1979) samples (uniformly sampled at random with replacement) and then multiple predictions are averaged. Applying the Bagging method in sparse regression has been shown to reduce estimation error when the sparsity level s is high for a specific sparsity pattern (Breiman, 1996).

However, individually solved predictors are not guaranteed to have the same support, and in the worst case, their average can be quite dense – its support size growing up to a multiple of the number of estimates. Bolasso was proposed to alleviate this problem (Bach, 2008a) by estimating the support from the intersection of all bootstrapped estimators. However, this strategy is very aggressive and during large noise cases, the supports of the estimators may not align and it recovers an extremely sparse solution.

In this paper, we propose to collaboratively enforce the row sparsity constraint among all predictors using the $\ell_{1,2}$ norm to resolve the support inconsistency issue in Bagging and avoid the overly aggressive Bolasso type of scheme. We name this algorithm JOBS (**J**oint-sparse **O**ptimization from **B**ootstrap **S**amples). The proposed method involves two key parameters: the bootstrap sample size L of random sampling with replacement from the original m measurements and the K number of those bootstrap vectors. JOBS improves the robustness of sparse recovery in challenging scenarios such as high noise, limited measurements, and in the presence of adversarial samples. A short summary of comparing JOBS to classical methods is in Table 1.

Methods	ℓ_1 min.	Bagging	JOBS
Robustness against large noise; against adversarial samples	baseline	better	better
	No	No	Yes
Sparsity	medium	dense	sparse
Optimal L	$= m$	small	smaller
Factor to ℓ_1 bound, with probability	1 1	$\sqrt{L/m} < 1$ $1 - e^{\mathcal{O}(K/L)}$	$\sqrt{L/m} < 1$ $1 - e^{\mathcal{O}(K/L^2)}$

Table 1: Comparison of different methods of sparse recovery. The factor in the last row is the term associated with measurement noise power $\|\boldsymbol{z}\|_2$.

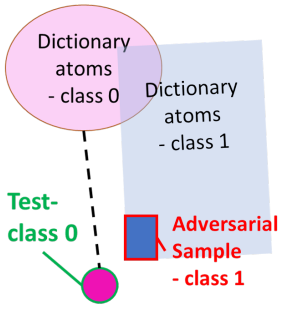

Figure 1: Adversarial sample is from a different class and is more similar to the test than dictionary atoms from the same class.

NOTATIONS: Let $\boldsymbol{A}$ denote the original sensing matrix of size $m \times n$. Let $\boldsymbol{y}$ represent the measurement vector. Let $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K$ be bootstrap samples, each containing L elements. For each bootstrapped sample $\mathcal{I}_j$, the corresponding bootstrapped sensing matrix $\boldsymbol{A}[\mathcal{I}_j]$ and bootstrapped measurements vector $\boldsymbol{y}[\mathcal{I}_j]$ are generated, where the operation $(\cdot)[\mathcal{I}]$ takes the rows of a matrix/ vector supported on $\mathcal{I}$. $\boldsymbol{x}_j$ is a feasible estimator for the j-th bootstrap sample. Concatenating K estimators $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_K$, we obtain the sparse-code matrix $\boldsymbol{X}$ of size $n \times K$. The row sparsity norm that we impose in the optimization is defined as the sum of the ℓ_2 norm of each row of this matrix: for $\boldsymbol{X}$, $\|\boldsymbol{X}\|_{1,2} = \sum(\|\boldsymbol{x}[1]^{\boldsymbol{T}}\|_2, \|\boldsymbol{x}[2]^{\boldsymbol{T}}\|_2, ..., \|\boldsymbol{x}[n]^{\boldsymbol{T}}\|_2)$.

The proposed method: JOBS consists of three steps. First, we generate K bootstrap samples: $\{\mathcal{I}_1, \mathcal{I}_2, .., \mathcal{I}_K\}$, each containing L indices. The bootstrapped data contains K pairs of sensing matrices measurements: $\{\boldsymbol{y}[\mathcal{I}_1], \boldsymbol{A}[\mathcal{I}_1]\}, \{\boldsymbol{y}[\mathcal{I}_2], \boldsymbol{A}[\mathcal{I}_2]\}...., \{\boldsymbol{y}[\mathcal{I}_K], \boldsymbol{A}[\mathcal{I}_K]\}$. Second, we solve the collaborative recovery on those sets. For parameter $\lambda_{L,K} > 0$ that balances the least squares fit and the joint sparsity penalty based on the choice of (L, K), the joint sparse optimization is:

$$\widehat{\boldsymbol{X}} = \min_{\boldsymbol{X}} \lambda_{L,K} \|\boldsymbol{X}\|_{1,2} + 0.5 \sum_{j=1}^{K} \|\boldsymbol{y}[\mathcal{I}_j] - \boldsymbol{A}[\mathcal{I}_j]\boldsymbol{x}_j\|_2^2. \tag{1}$$

The proposed form in $\mathbf{J_{12}^{\lambda}}$ is a special case of block (group) sparse recovery (Berg & Friedlander, 2008) and there are numerous optimization methods for solving them such as (Boyd et al., 2011; Baron et al., 2009; Heckel & Bolcskei, 2012; Sun et al., 2009; Bach, 2008b; Berg & Friedlander, 2008; Wright et al., 2009b; Deng et al., 2011). Finally, the JOBS solution is obtained by averaging the columns of the solution from (1):

$$\text{JOBS:} \quad \boldsymbol{x}^{\boldsymbol{J}} = \frac{1}{K}\sum_{j=1}^{K}\widehat{\boldsymbol{x}}_j. \tag{2}$$

2 Theoretical Results

2.1 Correctness of JOBS via Block Null Space Property (BNSP)

Block Null Space Property (BNSP), characterizes the exact recovery condition of our algorithm as a Necessary and sufficient condition of noiseless program (Gao et al., 2015). we established BNSP for JOBS and since it established characterizes the existence and uniqueness of the true noiseless JOBS solution, and then we prove the correctness of JOBS-noiseless defined in (14). Since the final estimate the average of the solution, the latter part of Theorem 2 implies that the JOBS solution is also optimal $\boldsymbol{x}^{\boldsymbol{J}} = \boldsymbol{x}^{\star}$. The detailed proof is shown in Appendix 8.

Definition 1 (BNSP for JOBS) *A set of bootstrapped sensing matrices* $\{\boldsymbol{A}[\mathcal{I}_1], \boldsymbol{A}[\mathcal{I}_2], ..., \boldsymbol{A}[\mathcal{I}_K]\}$ *satisfies BNSP of order* s *if* $\forall\ (\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_K) \in \text{Null}(\boldsymbol{A}[\mathcal{I}_1]) \times \text{Null}(\boldsymbol{A}[\mathcal{I}_2]) ... \times \text{Null}(\boldsymbol{A}[\mathcal{I}_K]) \backslash \{(\boldsymbol{0}, \boldsymbol{0}, ..., \boldsymbol{0})\}$, *such that for all* $\mathcal{S} : \mathcal{S} \subset \{1, 2, ..., n\}, card(\mathcal{S}) \leq s$, $\|\boldsymbol{V}[\mathcal{S}]\|_{1,2} < \|\boldsymbol{V}[\mathcal{S}^c]\|_{1,2}$.

Theorem 2 (Correctness of JOBS) *The noiseless JOBS program successfully recovers all the* $s-$*row sparse solution if and only if* $\{\boldsymbol{A}[\mathcal{I}_1], \boldsymbol{A}[\mathcal{I}_2], ..., \boldsymbol{A}[\mathcal{I}_K]\}$ *satisfies BNSP of the order of* s *described in Definition 1. The solution is of the form* $\boldsymbol{X}^{\star} = (\boldsymbol{x}^{\star}, \boldsymbol{x}^{\star}, ..., \boldsymbol{x}^{\star})$, *where* $\boldsymbol{x}^{\star}$ *is the unique true sparse solution. Then, the JOBS solution* $\boldsymbol{x}^{\boldsymbol{J}}$, *which is the average over columns of* $\boldsymbol{X}^{\star}$, *is* $\boldsymbol{x}^{\star}$.

2.2 Block Restricted Isometry Propertity (BRIP) of JOBS

Let the JOBS block diagonal matrix $\boldsymbol{A}^{\boldsymbol{J}} = \text{block\_diag}(\boldsymbol{A}[\mathcal{I}_1], \boldsymbol{A}[\mathcal{I}_2], ..., \boldsymbol{A}[\mathcal{I}_K])$, where block_diag denotes the operator that stacks matrices as a block diagonal matrices, and $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_n\}$ is the block partition of all indices of vectorized matrix $\boldsymbol{X} \in \mathbb{R}^{n \times K}$ that correspond to the row sparsity pattern. Let $\delta_{s|\mathcal{B}}$ denote row sparse Block Restrict Isometry Property (BRIP) constant of order s over a given block partition $\mathcal{B}$ and δ_s denote the standard RIP constant of order s. We have the following proposition for JOBS by using the induced vector norm form of eigenvalue function.

Proposition 3 (BRIP for JOBS) *For all* $s \leq n, s \in \mathbb{Z}^+$,

$$\delta_{s|\mathcal{B}}(\boldsymbol{A}^{\boldsymbol{J}}) = \max_{j=1,2,...,K} \delta_s(\boldsymbol{A}[\mathcal{I}_j]). \tag{3}$$

It is not surprising at all that the BRIP of JOBS depends on the worst case among all K bootstrapped matrices since a smaller RIP constant indicates better recovery ability. The proof of this proposition is elaborated in Appendix 10.

2.3 Noisy Recovery for JOBS

Next, we analyze the error bound for JOBS using BNSP and BRIP in the noisy case. Note that our theorems are based on *deterministic* sensing matrix, measurements and noise vectors: $\boldsymbol{A}, \boldsymbol{y}, \boldsymbol{z}$ and the *randomness* in our framework is introduced by *the bootstrap sampling* process.

From previous analysis, we have established that if the BRIP constant of order $2s$ is less than $\sqrt{2}-1$, it implies that $\{\boldsymbol{A}[\mathcal{I}_1], \boldsymbol{A}[\mathcal{I}_2], ..., \boldsymbol{A}[\mathcal{I}_K]\}$ satisfies BNSP of order s. Then, Theorem 2 establishes that the optimal solution to $\mathbf{J_{12}}$ the noiseless version of joint sparse optimization is the $s-$row sparse signal $\boldsymbol{X}^{\star}$ with every column being $\boldsymbol{x}^{\star}$. Similar to the bound in Theorem 2 in (Eldar & Mishali, 2009) ,the reconstruction error is determined by the $s-$block sparse approximation error and the noise level. The Hoeffding's tail bound is used to obtain the worst case performance for JOBS. The

following theorem states the performance bound for JOBS when the ground truth signal $\boldsymbol{x}^\star$ is exactly $s-$sparse.

The relationship to the upper bound of RIP constant is discussed in Section 2.5. In the more general case, when the sparsity level of $\boldsymbol{x}^\star$ possibly exceeds s, we use the we derived the following error bounded associated with measurements error and $s-$ sparse approximation error.

The error bound in Theorem 5 relates to $s-$sparse approximation error as well as the noise level, which is similar to ℓ_1 minimization and block sparse recovery bounds. JOBS also introduces a relaxation error bounded by $\|\boldsymbol{e}\|_2$, which is the distance of the true vector to its top $s-$ sparse approximation.

Theorem 4 (JOBS: error bound for $\|\boldsymbol{x}^\star\|_0 = s$) *Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{z}$, $\|\boldsymbol{z}\|_2 < \infty$. If there exists a constant related to parameters (L, K) such that, $\delta_{2s|\mathcal{B}}(\boldsymbol{A}^{\boldsymbol{J}}) \leq \delta_{L,K} < \sqrt{2} - 1$ and the true solution is exactly $s-$sparse, then for any $\tau > 0$, JOBS solution $\boldsymbol{x}^{\boldsymbol{J}}$ satisfies*

$$\mathbb{P}\left\{ \|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2 \leq \mathcal{C}_1(\delta_{L,K})(\sqrt{\frac{L}{m}}\|\boldsymbol{z}\|_2 + \tau) \right\} \geq 1 - \exp\frac{-2K\tau^4}{L\|\boldsymbol{z}\|_\infty^4}, \tag{4}$$

where $\mathcal{C}_1(\cdot)$ is the same non-decreasing functions of δ as in Theorem 1.3 in (Candes, 2008), which is reminded in Preliminary results session in Appendix 6.

Theorem 5 (JOBS: error bound for the general case) *Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{z}$, $\|\boldsymbol{z}\|_2 < \infty$. If there exists a constant related to parameters (L, K) such that, $\delta_{2s|\mathcal{B}}(\boldsymbol{A}^{\boldsymbol{J}}) \leq \delta_{L,K} < \sqrt{2} - 1$, then for any $\tau > 0$, JOBS solution $\boldsymbol{x}^{\boldsymbol{J}}$ satisfies*

$$\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2 \leq \mathcal{C}_0(\delta_{L,K})s^{-1/2}\|\boldsymbol{e}\|_1 + \mathcal{C}_1(\delta_{L,K})(\sqrt{\frac{L}{m}}\|\boldsymbol{z}\|_2 + \tau)\} \geq 1 - \exp\frac{-2K\tau^4}{L\|\boldsymbol{z}\|_\infty^4}. \tag{5}$$

where $\mathcal{C}_1(\cdot)$ is the same non-decreasing function of δ as in in Theorem 1.3 in (Candes, 2008); $\boldsymbol{e}$ is the s-sparse approximation error: $\boldsymbol{e} = \boldsymbol{x}^\star - \boldsymbol{x}_0$ with $\boldsymbol{x}_0$ containing the largest s components of the true solution $\boldsymbol{x}^\star$; and $\|\boldsymbol{A}\|_{\infty,1} = \max_{i=1,2,\ldots,m}(\|\boldsymbol{a}[i]^{\boldsymbol{T}}\|_1)$ denotes the largest ℓ_1-norm of all rows of $\boldsymbol{A}$.

We use Theorem 5 to explain the case when the number of measurements is low compared to the true sparsity level s. The trade-offs for a good choice of the bootstrap sample size L and the number of bootstrap samples K are discussed in Section 2.5.

2.4 Noisy Recovery for Bagging in Sparse Recovery

We now give the error bounds for employing the Bagging scheme in sparse recovery problems, in which the final estimate is the average over multiple estimates solved individually and independently from bootstrap samples.

Theorem 6 (Bagging: Error bound for $\|\boldsymbol{x}^\star\|_0 = s$) *Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{z}$, $\|\boldsymbol{z}\|_2 < \infty$. If there exists a constant related to parameters (L, K) such that, for all $j \in \{1, 2, ..., K\}$, $\delta_{2s}(\boldsymbol{A}[\mathcal{I}_j]) \leq \delta_{L,K} < \sqrt{2} - 1$, where $\boldsymbol{A}[\mathcal{I}_j]$ is the bootstrapped matrix. and let $\boldsymbol{x}^{\boldsymbol{B}}$ be the solution of Bagging, then, for any $\tau > 0$, $\boldsymbol{x}^{\boldsymbol{B}}$ satisfies*

$$\mathbb{P}\left\{ \|\boldsymbol{x}^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2 \leq \mathcal{C}_1(\delta_{L,K})(\sqrt{\frac{L}{m}}\|\boldsymbol{z}\|_2 + \tau) \right\} \geq 1 - \exp\frac{-2K\tau^4}{L^2\|\boldsymbol{z}\|_\infty^4}, \tag{6}$$

where $\mathcal{C}_1(\cdot)$ is the same non-decreasing function of δ as in Theorem 1.3 in (Candes, 2008).

Theorem 7 (Bagging: Error bound for the general case) *Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{z}$, $\|\boldsymbol{z}\|_2 < \infty$. If there exists a constant related to parameters (L, K) such that, for all $j \in \{1, 2, ..., K\}$, $\delta_{2s}(\boldsymbol{A}[\mathcal{I}_j]) \leq \delta_{L,K} < \sqrt{2} - 1$, and then, for any $\tau > 0$, the Bagging solution $\boldsymbol{x}^{\boldsymbol{B}}$ satisfies*

$$\mathbb{P}\left\{ \|\boldsymbol{x}^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2 \leq \mathcal{C}_0(\delta_{L,K})s^{-1/2}\|\boldsymbol{e}\|_1 + \mathcal{C}_1(\delta_{L,K})(\sqrt{\frac{L}{m}}\|\boldsymbol{z}\|_2 + \tau) \right\} \geq 1 - \exp\frac{-2K\tau^4}{(b')^2}. \tag{7}$$

where $\mathcal{C}_0(\cdot), \mathcal{C}_1(\cdot)$ are the same non-decreasing functions of δ as in Theorem 1.3 in (Candes, 2008), and $b' = (\mathcal{C}_0(\delta)\mathcal{C}_1{}^{-1}(\delta)s^{-1/2}\|\boldsymbol{e}\|_1 + \sqrt{L}\|\boldsymbol{z}\|_\infty)^2$.

Theorem 7 gives the performance bound for Bagging in general signal recovery without the $s-$sparse assumption, and it reduces to Theorem 6 when the $s-$sparse approximation error is zero, i.e., $\|\boldsymbol{e}\|_1 = 0$. Both Theorem 6 and 7 above show that increasing the number of estimates K improves the result by increasing the lower bound of the certainty for the same performance level.

JOBS vs Bagging bounds: The RIP condition for Bagging is the same as the RIP condition for JOBS, under the assumption that all bootstrapped matrices $\boldsymbol{A}[\mathcal{I}_j]$s are well-behaved for the worst case analysis. When $\|\boldsymbol{x}^\star\|_0 = s, \|\boldsymbol{e}\| = 0$, the bound in Bagging is worse than JOBS since the certainty for algorithm is at least $1 - \exp\frac{-2K\tau^4}{L^2\|\boldsymbol{z}\|_\infty^4}$, compared to the error bound $1 - \exp\frac{-2K\tau^4}{L\|\boldsymbol{z}\|_\infty^4}$ in JOBS. When $\|\boldsymbol{e}\| > 0$, we can derive (the right hand side) r.h.s. of Bagging (7) < the r.h.s. of Bagging in $s-$ sparse (6) < the r.h.s. of JOBS (5). With an L^2 term instead of L in the denominator, the bound is tighter for JOBS given the same L and K. This comparison shows that JOBS has a better theoretical worst-case performance bound for an s-sparse signal; recovery of a nearly s-sparse signal follows similar behavior.

2.5 Key Parameters (L, K) Selection from Theoretical Analysis

Concerning the sampling ratio L/m, two competing factors influence the optimal choice. In general, the BRIP constant decreases with increasing L; thus, more measurements leads to better recovery. Additionally, increasing L also results in smaller δ and $\mathcal{C}_1(\delta)$. However, larger L can also increase noise, evidenced by the second factor associated with the noise power term, $\sqrt{L/m}$. Thus, a moderate choice of the L/m ratio is best. Experimental results show best performance at $L/m \approx 0.4$. As m grows larger and problem becomes easier, the optimal L/m also increases.

As for the number of estimates K, increasing K weakly increases the BRIP constant, but not by a significant margin. In the sparse regression simulation, we find that increasing K in general does not degrade performances. Increasing K mainly reduces uncertainty in (5), which decays exponentially with K. The certainty can be written as $p(K) = 1 - \exp\{-\alpha K\}$, for some $\alpha > 0$. The growth rate of $dp(K)/dK$ is non-negative and decreasing with K. In short, although increasing K will in general improve the results, the improvement margin decreases as K gets larger. We validate this phenomenon in our simulation.

3 Experimental Results

Three experiments are done to investigate the property of JOBS: (i) on classic sparse reconstruction task on synthetic data, (ii) image reconstruction with presence of adversarial examples on real dataset (iii) standard image classification task on real dataset.

3.1 Sparse reconstruction from compressed measurements

In this section, we perform sparse recovery on a generic synthetic dataset to study the performance of the proposed algorithm. In our experiment, all entries of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ are i.i.d. samples from the standard normal distribution $\mathcal{N}(0, 1)$. The signal dimension $n = 200$, and various numbers of measurements from 50 to 150. The ground truth signals $\boldsymbol{x}^\star$ has its sparsity level set to $s = 50$. The location of each non-zeros entry is selected uniformly at random whereas its magnitude is sampled from the standard Gaussian distribution. For the noise processes $\boldsymbol{z}$, all entries are sampled i.i.d. from $\mathcal{N}(0, \sigma^2)$, with variance $\sigma^2 = 10^{-\text{SNR}/10}\|\boldsymbol{Ax}\|_2^2$, where SNR represents the Signal-to-Noise Ratio. In our experiment, we study three different noise levels: when SNR $= 0, 1$ and 2 dB. The same solver to solve sparse regression in all comparison methods: JOBS, Bagging, Bolasso, ℓ_1 minimization. Details is in Appendix 11.

We explore how two key parameters – the number of estimates K and the bootstrapping ratio L/m – affect sparse regression results. In our experiment, we vary $K = 30, 50, 100$ while setting the bootstrap ratio L/m from 0.1 to 1 with an increment of 0.1. We report the average recovered Signal to Noise Ratio (SNR) as the error measure to evaluate the recovery performance: $\text{SNR}(\widehat{\boldsymbol{x}}, \boldsymbol{x}^\star) = -10\log_{10}\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2^2/\|\boldsymbol{x}^\star\|_2^2$ (dB) averaged over 20 independent trials. For all algorithms, we vary the balancing parameter $\lambda_{L,K}$ at different values from .01 to 200 and then select the optimal value that gives the maximum averaged SNR over all trials at each (L, K).

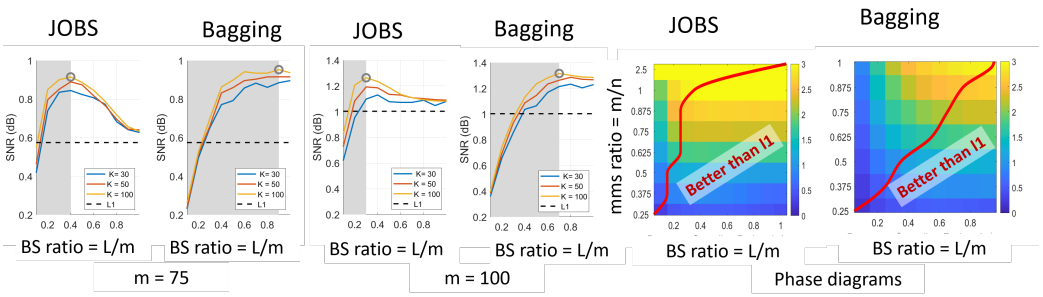

Figure 2: Recovery performance curves for JOBS and Bagging (with various L,K) versus ℓ_1 minimization. **Left 1-4:** The number of measurements are $m = 75, 100$ from left to right. **Left 5-6:** Phase diagrams of JOBS, Bagging. Noise level is set to SNR = 0 dB.

Performance of JOBS, ℓ_1 min., Bagging is illustrated in Figure 2 with different total numbers of measurements $m = 50, 150$. Note, for each condition with a particular choice of (L, K), the information available to JOBS, Bagging and Bolasso algorithms is identical and ℓ_1-minimization always has access to all m measurements. When the number of measurements m is limited, JOBS outperforms ℓ_1 minimization significantly. As m increases, the margin decreases. When the number of measurements is low (the sparsity level $s = 50$ and m is only $50 - 150$, which is between $1s - 3s$), and with very small bootstrap sampling ratio L/m (L/m is only $0.3 - 0.5$) JOBS and Bagging are quite robust and outperform all other algorithms using the same parameters (L, K). In addition, although JOBS and Bagging are similar in terms of the best performance limit, which are within 3% in our overall experiments. Bagging requires higher L/m ratios (typically ≥ 0.6) to achieve peak performance than JOBS. This is explored more in the next paragragh.

JOBS vs Bagging

• **JOBS Optimal Sampling Ratio is Consistently Smaller than That of Bagging.** Both JOBS and Bagging outperform the classical ℓ_1 minimization algorithm in the challenging case when the total number of measurements m is low. The peak performance of JOBS and Bagging are comparable (within 3%). Table 2 shows the optimal ratios for JOBS algorithm and for Bagging with the number of measurements m from $50 - 150$ and various SNR ratios SNR $= 0, 1, 2$ dB. The optimal bootstrap sampling ratio for JOBS is smaller than that for Bagging. With the same K as Bagging, JOBS achieves optimal performance with a much smaller vector size L compared to Bagging.

A smaller bootstrap size leads leads to a reduction of the algorithm complexity. With the ADMM implementation, the theoretical complexity levels for both Bagging and JOBS algorithms are the same for the same (L, K): $\mathcal{O}(n^2(L+n)K) + T\mathcal{O}(n^2K)$, where T is number of iterations. This result Since the optimal L is smaller, JOBS yields a smaller complexity than Bagging.

	SNR = 0		SNR = 1		SNR = 2	
m	JOBS	Bag.	JOBS	Bag.	JOBS	Bag.
50	0.5	0.6	0.6	0.8	0.5	0.8
75	0.4	0.9	0.4	1	0.4	0.7
100	0.3	0.7	0.3	1	0.4	1
150	0.4	1	0.5	1	0.5	1

Table 2: The Empirical Optimal Sampling Ratios L/m with Limited Measurements m. $K = 100$.

	Bootstrapping-based methods		
m	JOBS	Bagging	Bolasso
50	$89 \pm 3\%$	$91 \pm 2\%$	$0.03 \pm 0.1\%$
75	$78 \pm 4\%$	$82 \pm 5\%$	$0.20 \pm 0.4\%$
100	$71 \pm 4\%$	$91 \pm 2\%$	$0.25 \pm 0.3\%$
150	$47 \pm 6\%$	$87 \pm 5\%$	$3.6 \pm 1\%$

Table 3: The averaged sparsity ratios of recovered signals. The numerical threshold for being non-zero is 10^{-2}. SNR $= 0$ dB.

• **JOBS solutions are consistently sparser than Bagging solutions.** We check the sparsity of the reconstructed signals through the numeric sparsity ratio: the sparsity ratio for a reconstructed vector $\widehat{\boldsymbol{x}}$ is the ratio elements with whose magnitude higher than the threshold ($\tau > 0$) over all elements. From Table 3, JOBS generally produces sparser solutions than Bagging. It verifies our motivation to have more precise control over the sparsity level in JOBS algorithm than individually solved predictors such as Bagging, which are not guaranteed to have the same support on bootstrapped solutions.

3.2 IMAGE RECONSTRUCTION: JOBS IS ROBUST AGAINST ADVERSARIAL SAMPLES

This experiment verifies the robustness of JOBS in the presence of adversarial samples. A adversarial sample definition is illustrated as in Fig. 1. We evaluate two cases: **Case A - Adversarial sample from occlusion**: Test image, to be recovered, is of a woman with scarf. The dictionary does not include any images with a scarf from the same class, but it does include an adversarial sample from a different class: i.e. a man wearing a scarf. **Case B - Adversarial sample from misalignment**: Test image is of a woman. All dictionary atoms from the same class are rotated to various degrees. However, there is a picture of a man with the same alignment as the test image.

In both cases, we use a common face recognition dataset: the cropped AR dataset (Martinez & Kak, 2001), containing pre-aligned images taken in various controlled conditions. The dimensions of all images are 165×120 pixels. We took images from two people: one woman (with label W-001) and one man (with label M-001) as our dictionary and test signal to be reconstructed in both cases. We use simplified notation $W\#$ to indicate label $\#$ from W-001 and $M\#$ for pictures of the man. For each person, the same label corresponds to the same controlled condition.

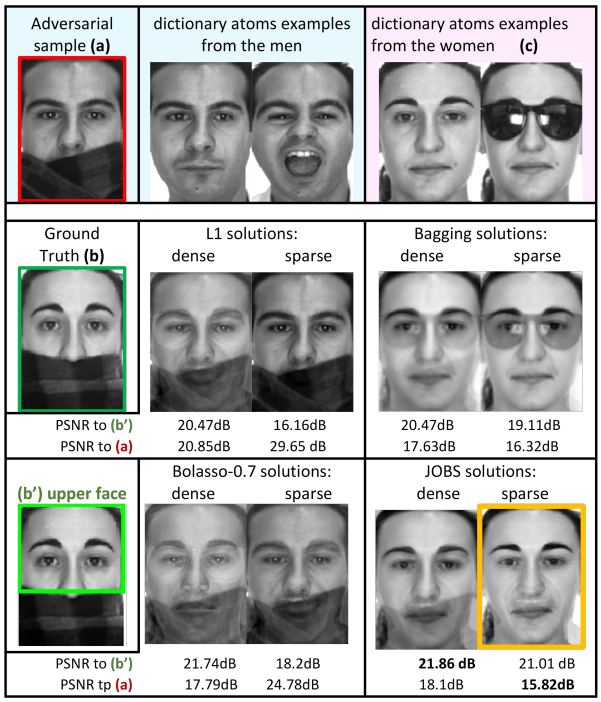

Figure 3: Reconstruction with adversarial sample with similar occlusion as test. **Top**: adversarial sample (a) and dictionary examples. **Mid-Bot.:** The test image (b) and reconstructions. ℓ_1 favors (a). Bagging solutions contain energy from (c), far from ground truth. Bolasso has visible components from (a). JOBS successfully avoids (a) as sparsity regularization increases (yellow box). Ground truth PSNRs calculated with respect to b′, the top 100 rows of (b).

Case A: The dictionary contains $W1 - W10$ from the woman and $M1 - M11$ from the man. The test image to be reconstructed is of the woman wearing the scarf: $W11$. $M11$ serves as an adversarial sample that may fool recovery methods due to a scarf occlusion very similar to that in the test image.

Parameters: The reconstruction is performed directly in the vectored image domain. As a variation for JOBS and Bagging, instead of picking random bootstrap samples, we pick 200 random 12×12 patches, to take advantage of the local robust features. We adopt a soft version of Bolasso: Bolasso-S to reduce the chance of zero solutions. The estimated support contains locations present in at least S replications, rather than requiring them to be in all K (Bach, 2008a). We take $S = 0.7$.

Bagging reconstruction is taken from the exact same set of measurements as JOBS. The sparsity regularization parameters for (dense, sparse) solutions for ℓ_1 are $(100, 5000)$; for Bagging are $(100, 2500)$; for Bolasso-0.7 are $(0.01, 1.2)$; and for JOBS are $(10^3, 10^4)$, respectively.

Performance from all four methods is shown in Figure 3. ℓ_1 minimization is fooled by the adversarial example and selects it at any sparsity level. Although Bagging does not suffer as much from the adversarial sample, its dense solution contains strong artifacts from an image with glasses as shown in Figure 3(c). Bolasso, like ℓ_1 min., is strongly influenced by the adversarial sample. In contrast, JOBS avoided the adversarial sample well: the component of the adversarial example in JOBS solution reduces with increasing sparsity regularization, and the scarf eventually becomes invisible.

Case B: The test is $W7$. The dictionary contains $W1 - W6$, each rotated 5° **incrementally** counterclockwise, as well as an adversarial sample $W7$. The test and adversarial images have the same alignment whereas all other dictionary atoms are misaligned.

Parameters: We pick 1200 random patches of dimension 3×3. The sparsity regularization parameters for ℓ_1, Bagging, Bolasso, and JOBS are 10^5, 600, 200 and 1500, respectively.

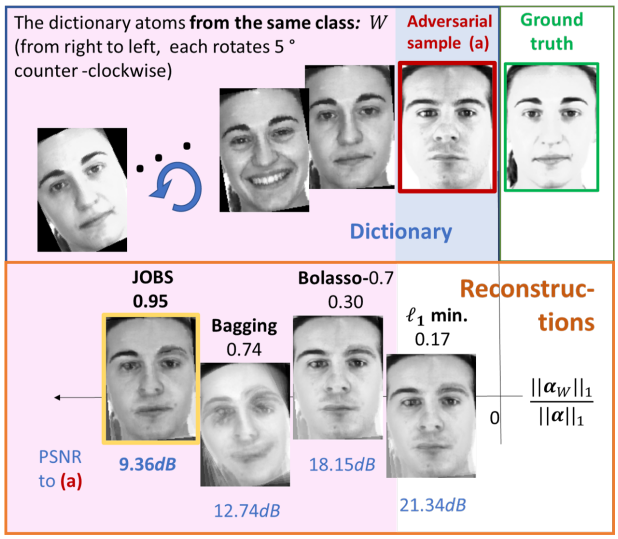

Figure 4: Reconstruction with adversarial sample with the same alignment condition as the test. **Top:** Adversarial sample (a) and examples in dictionary. **Bot.:** Reconstructions: ℓ_1 returns similarly to (a). Bagging solutions are too dense and therefore blurred. Bolasso contains large energy from (a). JOBS performs the best: a clear image with more than 90% from the correct class.

Performance of four algorithms is illustrated in Figure 4. Here we use two metrics: the ratio of reconstructed signal from atoms from women dictionary over all locations and the PSNR to adversarial sample (a). According to both measures, the order of robustness to (a) from weak to strong is ℓ_1 min., Bolasso, Bagging and JOBS. Although Bagging has a large component from the correct class (75%), it is too dense and the reconstructed picture is blurry due to different alignment conditions in the dictionary.

3.3 Image Classification

To confirm that improvements in regression directly lead to improvements in classification, we performed classification experiments on the same cropped AR dataset. We first use random projection (Gaussian matrix) for dimension reduction to generate $m = 50$ random features as measurements. Then we solve the sparse regression problem using all four algorithms all within a Sparse Representation-based Classification (SRC) framework proposed by Wright in (Wright et al., 2009a) to predict class label.

As shown in Table 4, classification based on sparse representations generated by JOBS shows a consistent improvement of 3% in classification accuracy over the baseline ℓ_1 minimizatioon. As with regression, the optimal bootstrapping ratio for JOBS at only 0.5 is lower than Bagging. The JOBS solution is also much sparser than Bagging's (threshold for being non-zero is 10^{-6}), similar to ℓ_1.

Table 4: Classification accuracy, optimal parameters, and sparsity comparison with various methods. Bagging is NOT directly used on classification but on bagged sparse code. Dataset is Cropped AR ($m = 50$), with training ratio 0.92.

Metrics	Baseline ℓ_1 min.	Bootstrapping-based methods JOBS	Bagging	Bolasso
Accuracy	0.855	**0.880**	0.855	0.790
Optimal $(L/m, K)$	(1,1)	(0.5,30)	(1,50)	(0.9,30)
Sparsity Ratio	3.6% ($\pm$0.5%)	2.7%($\pm$0.5%)	27% ($\pm$3%)	0.56% ($\pm$0.3%)

4 Conclusion

We propose a *collaborative* signal recovery framework named JOBS, motivated from powerful bootstrapping ideas in machine learning. JOBS improves the robustness of sparse recovery in challenging scenarios of noisy environments and/or limited measurements, and with the presence of adversarial samples. Below are highlights: (i) JOBS is particularly powerful when the number of measurements m is limited, outperforming ℓ_1 min. by a large margin. (ii) JOBS achieves desirable performances with relatively low bootstrap ratio L/m than Bagging and small number of bootstrapped observation vectors K. (iii) The optimal sampling ratio for collaborative JOBS is lower than that of independent Bagging while achieving similar results, resulting in a lower computation complexity. (iv) JOBS solutions are generally more sparse than Bagging's – a desirable property in sparse recovery. (v) JOBS is robust against adversarial samples.

REFERENCES

The birthday problem. http://www.math.uah.edu/stat/urn/Birthday.html. Creative Commons License.

Jose M Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In *Advances in Neural Information Processing Systems*, pp. 2270–2278, 2016.

F. R Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th Int. Conf. on Machine learning (ICML)*, pp. 33–40. ACM, 2008a.

F. R Bach. Consistency of the group lasso and multiple kernel learning. *The J. of Machine Learning Research*, 9:1179–1225, 2008b.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approx.*, 28(3):253–263, 2008.

D. Baron, M. F Duarte, M. B Wakin, S. Sarvotham, and R. G Baraniuk. Distributed compressive sensing. *arXiv preprint arXiv:0901.3403*, 2009.

E. Berg and M. P Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM J. on Scientific Computing*, 31(2):890–912, 2008. doi: 10.1137/080714488. URL http://link.aip.org/link/?SCE/31/890.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

P. L Bühlmann. Bagging, subagging and bragging for improving some prediction algorithms. In *Research Report*, volume 113. Seminar für Statistik, ETH Zurich, Switzerland, 2003.

P. L Bühlmann and B. Yu. Explaining bagging. In *Research Report*, volume 92. Seminar für Statistik, ETH Zurich, Switzerland, 2000.

E. J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

E. J Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Info. theory*, 52(2):489–509, 2006.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

E. Candess and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse prob.*, 23(3): 969, 2007.

S. Chen, D. L Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43 (1):129–159, 2001.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American mathematical society*, 22(1):211–231, 2009.

W. Deng, W. Yin, and Y. Zhang. Group sparse optimization by alternating direction method. In *Rice CAAM Report TR11-06*, pp. 88580R. International Society for Optics and Photonics, 2011.

D. L Donoho. Compressed sensing. *IEEE Trans. on Info. theory*, 52(4):1289–1306, 2006.

Julio Martin Duarte-Carvajalino and Guillermo Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Transactions on Image Processing*, 18(7):1395–1408, 2009.

B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Stat.*, 7(1):1–26, 1979.

Y. C Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Trans. on Info. Theory*, 55(11):5302–5316, 2009.

Y. Gao, J. Peng, and Y. Zhao. On the null space property of ℓ_q-minimization for in compressed sensing. *J. of Function Spaces*, 2015, 2015.

R. Heckel and H. Bolcskei. Joint sparsity with different measurement matrices. In *Proc. of 50th Annual Allerton Conf. on Communication, Control, and Computing, (ALLERTON)*, pp. 698–702. IEEE, 2012.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. of the American Statistical Association*, 58(301):13–30, 1963.

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

A. M Martinez and A. C Kak. PCA versus LDA. *IEEE Trans. on Pattern Anal. Mach. Intelligence*, 23(2):228–233, 2001.

A. F Mendelson, M. A Zuluaga, B. F Hutton, and S. Ourselin. What is the distribution of the number of unique original items in a bootstrap sample? *arXiv*, 2016.

L. Sun, J. Liu, J. Chen, and J. Ye. Efficient recovery of jointly sparse vectors. In *Advances in neural information processing systems (NeurIPs)*, pp. 1812–1820, 2009.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. of the Royal Stat. Society. Series B*, pp. 267–288, 1996.

I. Weiss. Limiting distributions in some occupancy problems. *The Annals of Mathematical Statistics*, 29(3):878–884, 1958.

J. Wright, A. Y Yang, A. Ganesh, S. S Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. on Pattern Anal. Mach. Intelligence*, 31(2):210–227, 2009a.

S. J Wright, R. D Nowak, and M. AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. on Sig. Proc.*, 57(7):2479–2493, 2009b.