

# High SNR Consistent Compressive Sensing

Sreejith Kallummil, Sheetal Kalyani  
 Department of Electrical Engineering  
 Indian Institute of Technology Madras,  
 Chennai, India 600036  
 {ee12d032,skalyani}@ee.iitm.ac.in

arXiv:1703.03596v1 [stat.ML] 10 Mar 2017

**Abstract**—High signal to noise ratio (SNR) consistency of model selection criteria in linear regression models has attracted a lot of attention recently. However, most of the existing literature on high SNR consistency deals with model order selection. Further, the limited literature available on the high SNR consistency of subset selection procedures (SSPs) is applicable to linear regression with full rank measurement matrices only. Hence, the performance of SSPs used in underdetermined linear models (a.k.a compressive sensing (CS) algorithms) at high SNR is largely unknown. This paper fills this gap by deriving necessary and sufficient conditions for the high SNR consistency of popular CS algorithms like  $l_0$ -minimization, basis pursuit denoising or LASSO, orthogonal matching pursuit and Dantzig selector. Necessary conditions analytically establish the high SNR inconsistency of CS algorithms when used with the tuning parameters discussed in literature. Novel tuning parameters with SNR adaptations are developed using the sufficient conditions and the choice of SNR adaptations are discussed analytically using convergence rate analysis. CS algorithms with the proposed tuning parameters are numerically shown to be high SNR consistent and outperform existing tuning parameters in the moderate to high SNR regime.

**Index Terms**—Compressive sensing, LASSO, Orthogonal matching pursuit, Dantzig selector, high SNR consistency.

## I. INTRODUCTION

Subset selection or variable selection in linear regression models is the identification of the support of regression vector  $\beta$ , i.e.,  $\mathcal{I} = \text{supp}(\beta) = \{j : \beta_j \neq 0\}$  in the regression model  $\mathbf{y} = \mathbf{X}\beta + \mathbf{w}$ . Here,  $\mathbf{X} \in \mathbb{R}^{n \times p}$  is a known design matrix with unit  $l_2$  norm columns,  $\mathbf{y} \in \mathbb{R}^n$  is the observed vector and  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$  is the additive white Gaussian noise with known variance  $\sigma^2$ . Let  $k^*$  denotes the number of non-zero entries in  $\beta$ . In this paper, we consider subset selection in underdetermined linear models, i.e.,  $\mathbf{X}$  with more columns than rows ( $n \leq p$ ). This problem studied under the compressive sensing (CS) paradigm is of fundamental importance in statistical signal processing, machine learning etc. Many compressive sensing (CS) algorithms with varying performance complexity trade-offs and optimality conditions are available [1]–[6] for this purpose. The performance of these CS algorithms are evaluated either in terms of mean square error (MSE) between  $\beta$  and the estimate  $\hat{\beta}$  returned by the CS algorithm [7] or the correctness with which the estimated support  $\hat{\mathcal{I}} = \text{supp}(\hat{\beta})$  matches the true support  $\mathcal{I}$  [8]. In this paper, we evaluate CS algorithms in terms of the probability of support recovery error defined by  $PE = \mathbb{P}(\hat{\mathcal{I}} \neq \mathcal{I})$ .

Traditionally, PE is evaluated in the large sample regime, i.e.,  $n \rightarrow \infty$  or  $(n, p) \rightarrow \infty$  [9]. In their landmark paper [10],

Ding and Kay demonstrated that subset selection procedures (SSPs) in overdetermined linear models that are large sample consistent (i.e.,  $PE \rightarrow 0$  as  $n \rightarrow \infty$ ) often performs poorly in a finite  $n$  and high signal to noise ratio (SNR) (i.e., small  $\sigma^2$ ) regime. This result generated great interest in the signal processing community on the behaviour of SSPs as  $\sigma^2 \rightarrow 0$ . Formally, a SSP is said to be high SNR consistent if its'  $PE \rightarrow 0$  as  $\sigma^2 \rightarrow 0$ . In this paper, we discuss the high SNR consistency of popular CS algorithms that are used for subset selection in underdetermined linear models. After presenting the mathematical notations, we elaborate on the existing literature on high SNR consistency and CS algorithms.

### A. Notations used in this paper.

$\text{col}(\mathbf{X})$  the column space of  $\mathbf{X}$ .  $\mathbf{X}^T$  is the transpose and  $\mathbf{X}^\dagger = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$  is the Moore-Penrose pseudo inverse of  $\mathbf{X}$  (if  $\mathbf{X}$  has full column rank).  $\mathbf{P}_{\mathbf{X}} = \mathbf{X} \mathbf{X}^\dagger$  is the projection matrix onto  $\text{col}(\mathbf{X})$ .  $\mathbf{I}_n$  represents an  $n \times n$  identity matrix and  $\mathbf{0}_n$  represents an  $n \times 1$  zero vector.  $\mathbf{X}_{\mathcal{J}}$  denotes the submatrix of  $\mathbf{X}$  formed using the columns indexed by  $\mathcal{J}$ .  $\mathbf{X}_{i,j}$  is the  $[i, j]^{\text{th}}$  entry of  $\mathbf{X}$ . If  $\mathbf{X}$  is clear from the context, we use the shorthand  $\mathbf{P}_{\mathcal{J}}$  for  $\mathbf{P}_{\mathbf{X}_{\mathcal{J}}}$ .  $\mathbf{a}_{\mathcal{J}}$  or  $\mathbf{a}(\mathcal{J})$  denotes the entries of  $\mathbf{a}$  indexed by  $\mathcal{J}$ .  $\mathcal{N}(\mathbf{u}, \mathbf{C})$  is a Gaussian vector with mean  $\mathbf{u}$  and covariance  $\mathbf{C}$ .  $\chi_j^2$  is a central chi square distribution with  $j$  degrees of freedom (d.o.f) and  $\chi_j^2(\lambda)$  is a non central chi square distribution with  $j$  d.o.f and non-centrality  $\lambda$ .  $\mathbf{a} \sim \mathbf{b}$  implies that  $\mathbf{a}$  and  $\mathbf{b}$  are identically distributed.  $|\cdot|$  denotes the absolute value for scalar arguments and cardinality for set arguments.  $\|\mathbf{a}\|_q = (\sum_j |\mathbf{a}_j|^q)^{\frac{1}{q}}$  for  $1 \leq q < \infty$  is the  $l_q$  norm,  $\|\mathbf{a}\|_\infty = \max_j |\mathbf{a}_j|$  is the  $l_\infty$  norm and  $\|\mathbf{a}\|_0 = |\text{supp}(\mathbf{a})|$  is the  $l_0$  quasi norm of  $\mathbf{a}$  respectively.  $\mathbf{a}$  is called  $k^*$ -sparse iff  $\|\mathbf{a}\|_0 = k^*$ .  $\|\mathbf{A}\|_{m,l} = \max_{\|\mathbf{x}\|_m=1} \|\mathbf{A}\mathbf{x}\|_l$  is the  $(m, l)^{\text{th}}$  matrix norm.  $[p]$  denotes the set  $\{1, \dots, p\}$ . For any two index sets  $\mathcal{J}_1$  and  $\mathcal{J}_2$ , the set difference  $\mathcal{J}_1/\mathcal{J}_2 = \{j \in \mathcal{J}_1 : j \notin \mathcal{J}_2\}$ .  $f(n) = o(g(n))$  iff  $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$ .

### B. Prior literature on high SNR consistency

Most of the existing literature related to high SNR consistency including the seminal work by Ding and Kay [10] are related to the model order selection (MOS) problem. MOS is a subset selection problem where  $\mathcal{I}$  is restricted to the form  $\mathcal{I} = [k^*]$ . Another interesting problem related to MOS is the estimation of smallest  $\tilde{k}$  such that  $\beta$  satisfies  $\beta_j = 0, \forall j > \tilde{k}$

and  $\beta_j$  can be zero or non-zero for  $j < \tilde{k}$ . In both these cases, the statistician is required to estimate the model order  $k^*$  or  $\tilde{k}$ . A number of MOS criteria like exponentially embedded family (EEF) [11], normalised maximum likelihood based minimum description length (NMDL) [12], g-prior based MDL (g-MDL) [13], forms of Bayesian Information criteria (BIC) [14], [15], sequentially normalised least squares (SNLS) [16] etc. are proved to high SNR consistent [10], [17]–[19]. All these MOS criteria can be formulated as the minimization of a penalised log likelihood

$$PLL(k) = \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}_k})\mathbf{y}\|_2^2 + h(k, \sigma^2)\sigma^2 \quad (1)$$

over the collection of subsets  $\{\mathcal{J}_k\}_{k=1}^p$ , where  $\mathcal{J}_k = [k]$  and  $h(k, \sigma^2)$  is a penalty function. Necessary and sufficient conditions (NSCs) for a MOS criterion to be high SNR consistent is derived in [17]. Applying MOS criteria to the general subset selection problem where  $\mathcal{I}$  can be any subset of  $[p]$  involves the minimization of  $PLL(\mathcal{J})$  over the entire  $2^p$  subsets  $\mathcal{J} \subseteq [p]$ . This approach though theoretically optimal is computationally intractable. Consequently a number of suboptimal but low complexity SSPs are developed. To the best of our knowledge, only two SSPs, both of which are based on the least squares (LS) estimate of  $\beta$  (i.e.,  $\hat{\beta}_{LS} = \mathbf{X}^\dagger \mathbf{y}$ ) are known to be high SNR consistent [8], [17].

### C. Contributions of this paper

The existing literature on high SNR consistency in linear regression is applicable only to regression models with full column rank design matrices. Hence, existing literature is not applicable to underdetermined linear models, i.e.,  $\mathbf{X}$  with  $n < p$ . Identifying the true support  $\mathcal{I}$  in an underdetermined linear model is an ill-posed problem unless certain structures are imposed on the regression vector  $\beta$  and design matrix  $\mathbf{X}$ . Throughout this paper, we assume that the regression vector  $\beta$  is sparse, i.e.,  $k^* = |\mathcal{I}| \ll p$  and  $k^* < n$ . The structure imposed on  $\mathbf{X}$  depends on the particular CS algorithm used.

This paper makes the following contributions to CS literature from the viewpoint of high SNR consistency. We first derive NSCs on the tuning parameter  $\Gamma_0$  such that the support estimate  $\hat{\mathcal{I}} = \text{supp}(\hat{\beta})$  delivered by

$$(l_0\text{-penalty}) : \hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \Gamma_0 \sigma^2 \|\mathbf{b}\|_0,$$

is high SNR consistent. It should be noted that optimization problem in  $l_0$ -penalty is NP-hard [20]. Hence, a number of suboptimal techniques broadly belonging to two classes, convex relaxation (CR) [2], [4] and greedy algorithms [3], [6] are developed in literature. We mainly consider two CR techniques in this paper, viz.,

$$(l_1\text{-penalty}) : \hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \sigma \Gamma_1 \|\mathbf{b}\|_1 \quad \text{and}$$

$$(l_1\text{-error}) : \hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{b}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2 \leq \sigma \Gamma_2.$$

$l_1$ -penalty and  $l_1$ -error are also known as basis pursuit de-noising (BPDN) or least absolute shrinkage and selection

operator (LASSO). We derive NSCs on  $\Gamma_1, \Gamma_2$  such that  $l_1$ -penalty and  $l_1$ -error are high SNR consistent. We also derive NSCs on the hyper parameter  $\Gamma_3$  of the popular CR technique Dantzig selector [2] given by

$$(DS) : \hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{b}\|_1 \quad \text{subject to} \quad \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{b})\|_\infty \leq \sigma \Gamma_3$$

for the special case of<sup>1</sup> orthonormal  $\mathbf{X}$ . Orthogonal matching pursuit (OMP) [3], [21]–[24] is a popular greedy algorithm with sound performance guarantees and low computational complexity in comparison with CR based SSPs. OMP is characterized by its' stopping condition (SC). We also derive high SNR consistent SCs for OMP.

Necessary conditions derived for  $l_0$ -penalty,  $l_1$ -penalty,  $l_1$ -error and DS analytically establish the high SNR inconsistency of these schemes with the values of  $\{\Gamma_k\}_{k=0}^3$  discussed in literature. High SNR inconsistency of OMP with popular SCs is numerically established. These inconsistencies are due to the absence of SNR adaptations in the tuning parameters. The sufficient conditions delivers a range of SNR adaptations for tuning parameters that will result in high SNR consistency. To compare various SNR adaptations, we derived simple bounds on the convergence rates of  $l_1$ -penalty. Extensive numerical simulations conducted on various subset selection scenarios demonstrate the potential of some of these SNR adaptations to significantly outperform existing tuning parameters in the moderate to high SNR regime. In addition to being a topic of theoretical importance, high SNR consistency of CS algorithms have tremendous practical value. A number of applications such as multi user detection [25], on-off random access [26], CS based single snapshot direction of arrival [27] etc. demands support recovery with very low values of PE in the moderate to high SNR regime. The high SNR consistent tuning parameters derived in this article can be applied directly for such applications in the moderate to high SNR regime.

### D. Organization of paper

Section II gives mathematical preliminaries. Section III discuss the high SNR consistency of  $l_0$ -penalty, Section IV discuss the consistency of CR techniques and Section V discuss the consistency of OMP. Section VI validates the analytical results through numerical simulations.

## II. MATHEMATICAL PRELIMINARIES

In this section, we present a brief overview of mathematical concepts from CS and probability theory used in this article.

### A. Qualifiers for design matrix $\mathbf{X}$ .

When  $n < p$ , the linear equation  $\mathbf{y} = \mathbf{X}\beta$  has infinitely many possible solutions. Hence the support recovery problem is ill-posed even in the noiseless case. To uniquely recover the

<sup>1</sup>In this article we consider a popular formulation of CS algorithms where the tuning parameters are explicitly scaled by  $\sigma$  or  $\sigma^2$ . Quite often  $\sigma$  or  $\sigma^2$  is included in the tuning parameter itself. For example,  $l_0$ -penalty may be written as  $\hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda_0 \|\mathbf{b}\|_0$ . Using the relation  $\lambda_0 = \sigma^2 \Gamma_0$ , the NSCs derived in terms of  $\Gamma_0$  can be easily restated in terms of  $\lambda_0$  also.

$k^*$ -sparse vector  $\beta$ , the measurement matrix  $\mathbf{X}$  has to satisfy certain well known regularity conditions.

**Definition 1:** The spark of a matrix  $\mathbf{X}$  ( $\text{spark}(\mathbf{X})$ ) is the smallest number of columns in  $\mathbf{X}$  that are linearly dependent.

Consider the following the optimization problem.

$$\hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{b}\|_0, \text{ subject to } \mathbf{y} = \mathbf{X}\mathbf{b}. \quad (2)$$

In words  $\hat{\beta}$  is the sparsest vector that solves the linear equation  $\mathbf{y} = \mathbf{X}\mathbf{b}$ . The following lemma relates the unique recovery of sparse vectors with  $\text{spark}(\mathbf{X})$  in the absence of noise.

**Lemma 1.** To uniquely recover all  $k^*$ -sparse vectors  $\beta$  using (1), it is necessary and sufficient that  $\text{spark}(\mathbf{X}) > 2k^*$  [1].

The optimization problem (2) cannot be solved in polynomial time. For polynomial complexity CS algorithms like DS,  $l_1$ -penalty,  $l_1$ -error, OMP etc.  $\text{spark}(\mathbf{X}) > 2k^*$  is not sufficient to guarantee unique recovery even in the noiseless case. A plethora of sufficient conditions including restricted isometry property (RIP) [1], [21], mutual incoherence condition (MIC) [4], [23], exact recovery condition (ERC) [3], [4] etc. are discussed in the literature. The high SNR analysis of CR techniques and OMP in this article uses ERC and MIC which are defined next.

**Definition 2:-** A matrix  $\mathbf{X}$  and a vector  $\beta$  with support  $\mathcal{I}$  is said to be satisfying ERC if the exact recovery coefficient  $\text{erc}(\mathbf{X}, \mathcal{I}) = \max_{j \notin \mathcal{I}} \|\mathbf{X}_{\mathcal{I}}^\dagger \mathbf{X}_j\|_1$  satisfies  $\text{erc}(\mathbf{X}, \mathcal{I}) < 1$ .

It is known that ERC is a sufficient and worst case necessary condition for accurately recovering  $\mathcal{I}$  from  $\mathbf{y} = \mathbf{X}\beta$  using OMP and the basis pursuit (BP) algorithm that solves

$$\hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{b}\|_1, \text{ subject to } \mathbf{y} = \mathbf{X}\mathbf{b} \quad (3)$$

in the noiseless case [3], [4]. ERC is also used to study the performance of  $l_1$ -penalty,  $l_1$ -error and OMP in noisy data [4], [23]. Since the ERC assumption involves the unknown support  $\mathcal{I}$ , it is impossible to check ERC in practice. Likewise, verifying the spark assumption is computationally intractable. Hence, the MIC, an assumption which can be easily verified is popular in CS literature [23].

**Definition 3:-** A  $k^*$ -sparse vector  $\beta$  satisfies MIC, iff the mutual coherence  $\mu_{\mathbf{X}} = \max_{i \neq j} |\mathbf{X}_i^T \mathbf{X}_j|$  satisfies  $\mu_{\mathbf{X}} < \frac{1}{2k^* - 1}$ .

If  $\mu_{\mathbf{X}} < \frac{1}{2k^* - 1}$ , then ERC is satisfied for all  $k^*$ -sparse vector  $\beta$ , i.e.,  $\text{erc}(\mathbf{X}, \mathcal{I}) < 1$  [3]. Likewise, MIC guarantees that  $\text{spark}(\mathbf{X}) > 2k^*$  [3]. Since, MIC implies both ERC and spark assumption, the analysis conducted based on ERC and spark are automatically applicable to problems satisfying MIC.

*Remark 1.* The number of measurements  $n$  is an important factor in deciding the properties of  $\mathbf{X}$  like spark,  $\mu_{\mathbf{X}}$  etc. In this paper, we will not explicitly quantify  $n$ , however by stating conditions on  $\text{spark}(\mathbf{X})$ ,  $\mu_{\mathbf{X}}$ , ERC etc. we implicitly assume that  $n$  is sufficiently large enough to satisfy these conditions.

### B. Standard Convergence concepts [Chapter 4, [28]].

A collection of random variables (R.Vs)  $X_{\sigma^2}$  converges in probability (C.I.P) to a R.V  $Y$ , i.e.,  $X_{\sigma^2} \xrightarrow{P} Y$  as  $\sigma^2 \rightarrow 0$  iff

$\forall \epsilon > 0, \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(|X_{\sigma^2} - Y| > \epsilon) = 0$ . A R.V  $X$  is B.I.P iff it is finite almost everywhere, i.e., for any  $\epsilon > 0, \exists R_\epsilon < \infty$  such that  $\mathbb{P}(|X| > R_\epsilon) < \epsilon$ . For an event  $A, \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(A) = 0$  iff for each  $\epsilon > 0, \exists \sigma_*^2(\epsilon) > 0$  such that  $\mathbb{P}(A) \leq \epsilon, \forall \sigma^2 < \sigma_*^2(\epsilon)$ . Next we describe the relationship between projection matrices and  $\chi^2$  R.Vs [17].

**Lemma 2.** Let  $\mathbf{P}$  be an arbitrary  $n \times n$  projection matrix with rank  $j$ . Then for any  $\mathbf{z} \sim \mathcal{N}(\mathbf{u}, \sigma^2 \mathbf{I}_n)$ ,  $\frac{\|\mathbf{P}\mathbf{z}\|_2^2}{\sigma^2} \sim \chi_j^2(\frac{\|\mathbf{P}\mathbf{u}\|_2^2}{\sigma^2})$  and  $\frac{\|(\mathbf{I}_n - \mathbf{P})\mathbf{z}\|_2^2}{\sigma^2} \sim \chi_{n-j}^2(\frac{\|(\mathbf{I}_n - \mathbf{P})\mathbf{u}\|_2^2}{\sigma^2})$ . Consider the two full rank sub matrices  $\mathbf{X}_{\mathcal{J}_1}$  and  $\mathbf{X}_{\mathcal{J}_2}$  formed by columns of  $\mathbf{X}$  indexed by  $\mathcal{J}_1 \subset \mathcal{J}_2$ . Let  $\mathbf{P}_{\mathcal{J}_1}$  and  $\mathbf{P}_{\mathcal{J}_2}$  represent the projection matrices onto the column space of  $\mathbf{X}_{\mathcal{J}_1}$  and  $\mathbf{X}_{\mathcal{J}_2}$  respectively. Then for any R.V  $\mathbf{z} \sim \mathcal{N}(\mathbf{u}, \sigma^2 \mathbf{I}_n)$ ,  $\frac{\|(\mathbf{P}_{\mathcal{J}_1} - \mathbf{P}_{\mathcal{J}_2})\mathbf{z}\|_2^2}{\sigma^2} \sim \chi_{|\mathcal{J}_2| - |\mathcal{J}_1|}^2(\frac{\|(\mathbf{P}_{\mathcal{J}_1} - \mathbf{P}_{\mathcal{J}_2})\mathbf{u}\|_2^2}{\sigma^2})$ .

Next we state a frequently used convergence result [17].

**Lemma 3.** Let  $z \sim \chi_j^2(\frac{\lambda}{\sigma^2})$ , where  $\lambda > 0$  is a constant w.r.t  $\sigma^2$ . Then  $\sigma^2 z \xrightarrow{P} \lambda$  as  $\sigma^2 \rightarrow 0$ .

### C. High SNR consistency: Definition

The high SNR consistency results available in literature [10], [17] deals with full rank linear regression models. Since, uniqueness issues are absent when  $\text{rank}(\mathbf{X}) = p$ , this definition of high SNR consistency demands that  $PE \rightarrow 0$  as  $\sigma^2 \rightarrow 0$  for every signal  $\beta \in \mathbb{R}^p$ . In this article, we relax this definition to account for the uniqueness issues present in regression models with  $n < p$  using the concept of regression class. A regression class  $\mathcal{C}$  is defined as the collection of matrix signal pairs  $(\mathbf{X}, \beta)$  where perfect recovery is possible for a particular algorithm under noiseless conditions. For  $l_0$ -penalty,  $\mathcal{C}_1 = \{(\mathbf{X}, \beta) : \text{spark}(\mathbf{X}) > 2|\text{supp}(\beta)|\}$  is a regression class. Similarly,  $\mathcal{C}_2 = \{(\mathbf{X}, \beta) : \mu_{\mathbf{X}} \leq \frac{1}{2|\text{supp}(\beta)| - 1}\}$  and  $\mathcal{C}_3 = \{(\mathbf{X}, \beta) : \text{erc}(\mathbf{X}, \text{supp}(\beta)) < 1\}$  forms regression classes for  $l_1$ -penalty,  $l_1$ -error and OMP. We now formally define high SNR consistency in underdetermined regression models.

**Definition 4:-** A SSP is said to be high SNR consistent for a regression class  $\mathcal{C}$  if  $PE = \mathbb{P}(\hat{\mathcal{I}} \neq \mathcal{I})$  converges to zero as  $\sigma^2 \rightarrow 0$  for every matrix vector pair  $(\mathbf{X}, \beta) \in \mathcal{C}$ .

In words, a SSP is high SNR consistent if it can deliver a  $PE$  arbitrarily close to zero by decreasing the noise variance  $\sigma^2$ . Even though every signal in a regression class can be perfectly recovered under noiseless conditions ( $\sigma^2 = 0$ ), to achieve a near perfect recovery at high SNR (i.e.,  $\sigma^2 \neq 0$ , but close to zero), the tuning parameters for the SSPs need to be selected appropriately. In the following sections, we discuss the conditions on the tuning parameters such that the support can be recovered with arbitrary precision as  $\sigma^2$  decreases.

### III. HIGH SNR CONSISTENCY OF $l_0$ -PENALTY BASED SSP.

In this section, we describe the high SNR behaviour of  $\hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \Gamma_0 \sigma^2 \|\mathbf{b}\|_0$  and  $\hat{\mathcal{I}} = \text{supp}(\hat{\beta})$ , where the tuning parameter  $\Gamma_0$  is a deterministic positive



quantity. The values of  $\Gamma_0$  discussed in the literature includes the Akaike information criteria (AIC) with  $\Gamma_0 = 2$ , minimum description length (MDL) or Bayesian information criteria (BIC) with  $\Gamma_0 = \log(n)$ , risk inflation criteria (RIC) of Foster and George (RIC-FG) with  $\Gamma_0 = 2 \log(p)$  [29], RIC of Zhang and Shen (RIC-ZS) with  $\Gamma_0 = 2 \log(p) + 2 \log(\log(p))$  [30], extended Bayesian information criterion (EBIC) with  $\Gamma_0 = \log(n) + \frac{2\gamma}{\|\mathbf{b}\|_0} \log(\binom{p}{\|\mathbf{b}\|_0})$  [31] etc. The hyper parameter  $\gamma$  in EBIC is a user defined parameter. Under a set of regularity conditions on the matrix  $\mathbf{X}$  and  $\beta$ , it was shown that  $l_0$ -penalty is large sample consistent if,  $\Gamma_0 = o(n^{c_2 - c_1})$ ,  $k^* \log(p) = o(n^{c_2 - c_1})$  and  $\Gamma_0 - 2 \log(p) - \log(\log(p)) \rightarrow \infty$  as  $n \rightarrow \infty$ . Here,  $c_1$  and  $c_2$  are parameters depending on the regularity conditions [32]. This result hold true for  $(n, p, k^*) \rightarrow \infty$  and  $n < p$  or  $n \ll p$ . Note that these tuning parameters are derived based on the large sample behaviour of  $l_0$ -penalty. The conditions for high SNR consistency of  $l_0$ -penalty are not discussed in the literature to the best of our knowledge. Next we state and prove the sufficient conditions for the high SNR consistency of  $l_0$ -penalty.

**Theorem 1.** Consider a matrix  $\mathbf{X}$  which satisfies  $\text{spark}(\mathbf{X}) > 2k^*$ . Then for any  $k^*$ -sparse signal  $\beta$ ,  $l_0$ -penalty is high SNR consistent if  $\lim_{\sigma^2 \rightarrow 0} \Gamma_0 = \infty$  and  $\lim_{\sigma^2 \rightarrow 0} \sigma^2 \Gamma_0 = 0$ .

*Proof.* The optimization problem in  $l_0$ -penalty can be stated more explicitly as  $\hat{\mathcal{I}} = \arg \min_{\mathcal{J} \subseteq [p]} L(\mathcal{J})$ , where  $L(\mathcal{J}) = \min_{\mathbf{b}: \text{supp}(\mathbf{b})=\mathcal{J}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \Gamma_0 \sigma^2 |\mathcal{J}|$ . When  $\mathbf{X}_{\mathcal{J}}$  has full rank,

the solution to  $\min_{\mathbf{b}: \text{supp}(\mathbf{b})=\mathcal{J}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 = \min_{\mathbf{a} \in \mathbb{R}^{|\mathcal{J}|}} \|\mathbf{y} - \mathbf{X}_{\mathcal{J}}\mathbf{a}\|_2^2$  is unique and equal to  $\hat{\mathbf{a}} = (\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}})^{-1} \mathbf{X}_{\mathcal{J}}^T \mathbf{y}$ . In this case,  $\mathbf{X}_{\mathcal{J}} \hat{\mathbf{a}} = \mathbf{P}_{\mathcal{J}} \mathbf{y}$  and  $\min_{\mathbf{b}: \text{supp}(\mathbf{b})=\mathcal{J}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2$  is equal to  $\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{y}\|_2^2$ . Here  $\mathbf{P}_{\mathcal{J}} = \mathbf{X}_{\mathcal{J}} (\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}})^{-1} \mathbf{X}_{\mathcal{J}}^T$  is a projection matrix of rank  $|\mathcal{J}| = \text{rank}(\mathbf{X}_{\mathcal{J}})$ . When  $\mathbf{X}_{\mathcal{J}}$  is rank deficient, the solution to  $\min_{\mathbf{b}: \text{supp}(\mathbf{b})=\mathcal{J}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 = \min_{\mathbf{a} \in \mathbb{R}^{|\mathcal{J}|}} \|\mathbf{y} - \mathbf{X}_{\mathcal{J}}\mathbf{a}\|_2^2$  can be any one of the infinitely many

vectors  $\hat{\mathbf{a}}$  that solves  $\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}} \hat{\mathbf{a}} = \mathbf{X}_{\mathcal{J}}^T \mathbf{y}$ . A typical solution is denoted by  $\hat{\mathbf{a}} = (\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}})^- \mathbf{X}_{\mathcal{J}}^T \mathbf{y}$ , where  $(\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}})^-$  is called the generalized inverse of  $\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}}$  [33]. The matrix  $\mathbf{X}_{\mathcal{J}} (\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}})^- \mathbf{X}_{\mathcal{J}}^T$  satisfies all the properties of a projection matrix of rank  $(\mathcal{X}_{\mathcal{J}})$ . We denotes this matrix by  $\mathbf{P}_{\mathcal{J}}$  itself with a caveat that  $\text{rank}(\mathbf{P}_{\mathcal{J}}) = \text{rank}(\mathbf{X}_{\mathcal{J}}) < |\mathcal{J}|$ . With this convention, when  $\mathbf{X}_{\mathcal{J}}$  is rank deficient,  $\min_{\mathbf{b}: \text{supp}(\mathbf{b})=\mathcal{J}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 = \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{y}\|_2^2$ . Hence,  $l_0$ -penalty can be reformulated as

$$\hat{\mathcal{I}} = \arg \min_{\mathcal{J} \subseteq [p]} L(\mathcal{J}) = \arg \min_{\mathcal{J} \subseteq [p]} \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{y}\|_2^2 + \sigma^2 \Gamma_0 |\mathcal{J}|. \quad (4)$$

Define the error event  $\mathcal{E} = \{\hat{\mathcal{I}} \neq \mathcal{I}\} = \{\exists \mathcal{J} \in [p] : L(\mathcal{J}) \leq L(\mathcal{I})\}$ . Applying union bound to  $PE = \mathbb{P}(\mathcal{E})$  gives

$$\begin{aligned} PE &\leq \sum_{\mathcal{J} \in [p]} \mathbb{P}(L(\mathcal{J}) \leq L(\mathcal{I})). \\ &= \overbrace{\sum_{\mathcal{J} \in \mathcal{H}_1} \mathbb{P}(L(\mathcal{J}) \leq L(\mathcal{I}))}^{P_1} + \overbrace{\sum_{\mathcal{J} \in \mathcal{H}_2} \mathbb{P}(L(\mathcal{J}) \leq L(\mathcal{I}))}^{P_2}. \end{aligned} \quad (5)$$

where  $\mathcal{H}_1 = \{\mathcal{J} \in [p] : (\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{X} \beta \neq \mathbf{0}_n\}$  and  $\mathcal{H}_2 = \{\mathcal{J} \in [p] : (\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{X} \beta = \mathbf{0}_n\}$ . In words,  $\mathcal{H}_1$  represent the subsets  $\mathcal{J} \subseteq [p]$  such that the  $\text{col}(\mathbf{X}_{\mathcal{J}})$  does not cover the signal subspace  $\text{col}(\mathbf{X}_{\mathcal{I}})$ . For  $\mathcal{I} = \{1, 2\}$ , assuming that the columns  $\mathbf{X}_1$ ,  $\mathbf{X}_2$  and  $\mathbf{X}_3$  are linearly independent, the subsets  $\mathcal{J} = \{1\}$ ,  $\mathcal{J} = \{3\}$ ,  $\mathcal{J} = \{1, 3\}$  etc. belongs to  $\mathcal{H}_1$ . Similarly,  $\mathcal{H}_2$  represents the subsets  $\mathcal{J} \subseteq [p]$  such that the  $\text{col}(\mathbf{X}_{\mathcal{J}})$  cover the signal subspace  $\text{col}(\mathbf{X}_{\mathcal{I}})$ . For  $\mathcal{I} = \{1, 2\}$ ,  $\mathcal{J} = \{1, 2, 3\}$ ,  $\mathcal{J} = \{1, 2, 3, 4\}$  etc. will belong to  $\mathcal{H}_2$ . We consider both these summations separately.

**Case 1**  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{X} \beta \neq \mathbf{0}_n$ : In this case, it can happen that  $|\mathcal{J}| > k^*$ ,  $|\mathcal{J}| = k^*$  or  $|\mathcal{J}| < k^*$ . Since  $\mathcal{I} = \text{supp}(\beta)$ ,  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}}) \mathbf{X} \beta = \mathbf{0}_n$ . Thus, by Lemma 2,  $A_1 = \frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}}) \mathbf{y}\|_2^2}{\sigma^2} \sim \chi_{n-k^*}^2$ . Likewise,  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{X} \beta \neq \mathbf{0}_n$  implies that  $A_2 = \frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{y}\|_2^2}{\sigma^2} \sim \chi_{n-\text{rank}(\mathbf{X}_{\mathcal{J}})}^2 \left(\frac{\lambda_{\mathcal{J}}}{\sigma^2}\right)$ , where  $\lambda_{\mathcal{J}} = \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{X} \beta\|_2^2 > 0$ . Hence,

$$\begin{aligned} \mathbb{P}(\mathcal{E}_{\mathcal{J}}) &= \mathbb{P}(L(\mathcal{J}) < L(\mathcal{I})) \\ &= \mathbb{P}((A_2 - A_1) \sigma^2 + \Gamma_0 \sigma^2 (|\mathcal{J}| - k^*) < 0). \end{aligned} \quad (6)$$

Since,  $A_1 \sim \chi_{n-k^*}^2$  is a B.I.P R.V,  $A_1 \sigma^2 \xrightarrow{P} 0$  as  $\sigma^2 \rightarrow 0$ . By Lemma 3,  $\sigma^2 A_2 \xrightarrow{P} \lambda_{\mathcal{J}} > 0$  as  $\sigma^2 \rightarrow 0$ . By the hypothesis of Theorem 1,  $\Gamma_0 \sigma^2 (|\mathcal{J}| - k^*) \rightarrow 0$  as  $\sigma^2 \rightarrow 0$ . This implies that  $(A_2 - A_1) \sigma^2 + \Gamma_0 \sigma^2 (|\mathcal{J}| - k^*) \xrightarrow{P} \lambda_{\mathcal{J}} > 0$ . Now, by the definition of C.I.P, for any  $\epsilon > 0$ ,  $\exists \sigma_{\mathcal{J}}^2 > 0$  such that  $\mathbb{P}\left(|(A_2 - A_1) \sigma^2 + \Gamma_0 \sigma^2 (|\mathcal{J}| - k^*) - \lambda_{\mathcal{J}}| > \frac{\lambda_{\mathcal{J}}}{2}\right) < \epsilon$ , for all  $\sigma^2 < \sigma_{\mathcal{J}}^2$ . This implies that

$$\mathbb{P}(\mathcal{E}_{\mathcal{J}}) \leq \mathbb{P}\left((A_2 - A_1) \sigma^2 + \Gamma_0 \sigma^2 (|\mathcal{J}| - k^*) < \frac{\lambda_{\mathcal{J}}}{2}\right) \leq \epsilon, \quad (7)$$

$\forall \sigma^2 < \sigma_{\mathcal{J}}^2$ . Thus,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_{\mathcal{J}}) = 0$ ,  $\forall \mathcal{J} \in \mathcal{H}_1$ . This together with  $|\mathcal{H}_1| < \infty$  implies that  $\lim_{\sigma^2 \rightarrow 0} P_1 = 0$ .

**Case 2**  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{X} \beta = \mathbf{0}_n$ :  $\text{spark}(\mathbf{X}) > 2k^*$  implies that  $\beta$  is the sparsest solution to the equation  $\mathbf{X} \mathbf{b} = \mathbf{X} \beta$ . Hence,  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{X} \beta = \mathbf{0}_n$  implies that  $|\mathcal{J}| > k^*$ . Since,  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{X} \beta = \mathbf{0}_n$ ,  $A_2 = \frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{y}\|_2^2}{\sigma^2} \sim \chi_{n-\text{rank}(\mathbf{X}_{\mathcal{J}})}^2$ . Thus  $\mathbb{P}(\mathcal{E}_{\mathcal{J}})$  becomes

$$\mathbb{P}(\mathcal{E}_{\mathcal{J}}) = \mathbb{P}(L(\mathcal{J}) < L(\mathcal{I})) = \mathbb{P}((A_1 - A_2) > \Gamma_0 (|\mathcal{J}| - k^*)). \quad (8)$$

Note that both  $A_1$  and  $A_2$  are B.I.P R.Vs with distribution independent of  $\sigma^2$  and so is  $A_1 - A_2$ . Thus,  $\exists t_{\epsilon} < \infty$  independent of  $\sigma^2$  such that  $\mathbb{P}(A_1 - A_2 > t_{\epsilon}) < \epsilon$ . Since,  $|\mathcal{J}| > k^*$ , by the hypothesis of Theorem 1,  $\Gamma_0 (|\mathcal{J}| - k^*) \rightarrow \infty$  as  $\sigma^2 \rightarrow 0$ . Thus,  $\exists \sigma_{\mathcal{J}}^2 > 0$ , such that  $\Gamma_0 (|\mathcal{J}| - k^*) > t_{\epsilon}$ ,  $\forall \sigma^2 < \sigma_{\mathcal{J}}^2$ . Combining, we get  $\mathbb{P}(\mathcal{E}_{\mathcal{J}}) < \epsilon$ ,  $\forall \sigma^2 < \sigma_{\mathcal{J}}^2$ . Thus,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_{\mathcal{J}}) = 0$ ,  $\forall \mathcal{J} \in \mathcal{H}_2$ . This together with  $|\mathcal{H}_2| < \infty$  implies that  $\lim_{\sigma^2 \rightarrow 0} P_2 = 0$ . Thus, under the hypothesis of Theorem 1,  $l_0$ -penalty is high SNR consistent. ■

*Remark 2.* Theorem 1 details a range of SNR adaptations on  $\Gamma_0$  such that  $l_0$ -penalty is high SNR consistent. However, different SNR adaptations satisfying Theorem 1 leads to

different convergence rates of  $PE$ . The proof of Theorem 1 reveals that  $P_1$  is related to the probability of underestimation and  $P_2$  is related to the probability of overestimation in MOS problems detailed in [17]. To summarise,  $\Gamma_0$  with faster rate of increase to  $\infty$  will have lower values of  $P_2$  and higher values of  $P_1$  and vice versa.

#### A. High SNR consistency of $l_0$ -penalty: Necessary conditions

The SNR adaptations required by Theorem 1 are in sharp contrast to the  $\sigma^2$  independent values of  $\Gamma_0$  discussed in literature. The following theorem proves that  $l_0$ -penalty with  $\sigma^2$  independent values of  $\Gamma_0$  are inconsistent at high SNR.

**Theorem 2.** Consider a matrix  $\mathbf{X}$  with  $\text{spark}(\mathbf{X}) > 2k^*$ . Then for any  $k^*$ -sparse vector  $\beta$ ,  $l_0$ -penalty is high SNR consistent only if  $\lim_{\sigma^2 \rightarrow 0} \Gamma_0 = \infty$ .

*Proof.* Define  $\mathcal{J} = \mathcal{I} \cup i$ , where  $i \notin \mathcal{I}$ . Note that  $|\mathcal{J}| = k^* + 1 \leq 2k^*$ ,  $\forall k^* \geq 1$  and  $|\mathcal{J}| = 1$  if  $k^* = 0$ . Further for any matrix  $\mathbf{X}$ ,  $\text{spark}(\mathbf{X}) \geq 2$ . Hence,  $\text{spark}(\mathbf{X}) > 2k^*$  implies that  $\mathbf{X}_{\mathcal{J}}$  has full rank for  $k^* \geq 0$ . This together with  $\mathcal{I} \subset \mathcal{J}$  implies that  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}})\mathbf{X}\beta = \mathbf{0}_n$ . Expanding  $L(\mathcal{J}) = \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}})\mathbf{y}\|_2^2 + \sigma^2\Gamma_0|\mathcal{J}|$  and applying Lemma 2, we have

$$PE \geq \mathbb{P}(L(\mathcal{J}) < L(\mathcal{I})) \geq \mathbb{P}(A > \Gamma_0), \quad (9)$$

where  $A = \frac{\mathbf{y}^T(\mathbf{P}_{\mathcal{J}} - \mathbf{P}_{\mathcal{I}})\mathbf{y}}{\sigma^2} \sim \chi_1^2$ ,  $\forall \sigma^2 > 0$ .  $A \sim \chi_1^2$  implies that  $A = Z^2$ , where  $Z \sim \mathcal{N}(0, 1)$ . Thus,  $PE \geq \mathbb{P}(A > \Gamma_0) = \mathbb{P}(|Z| > \sqrt{\Gamma_0}) = 2Q(\sqrt{\Gamma_0})$ ,  $\forall \sigma^2 > 0$ . Here  $Q(x) = \frac{1}{\sqrt{2\pi}} \int_{t=x}^{\infty} \exp(-\frac{t^2}{2}) dt$  is the complementary cumulative distribution function of a  $\mathcal{N}(0, 1)$  R.V. Hence,  $l_0$ -penalty is high SNR consistent only if  $\lim_{\sigma^2 \rightarrow 0} \Gamma_0 = \infty$ . ■

*Remark 3.* It follows directly from the proof of Theorem 2 that PE of  $l_0$ -penalty with SNR independent  $\Gamma_0$  like BIC, AIC etc. satisfy  $PE \geq 2Q(\sqrt{\Gamma_0})$ ,  $\forall \sigma^2 > 0$ . For RIC-FG with  $\Gamma_0 = 2\log(p)$ , the lower bound  $2Q(\sqrt{\Gamma_0})$  will be less than 0.01 only for  $p \geq 28$  and  $2Q(\sqrt{\Gamma_0}) \leq 0.001$  only for  $p \geq 225$ . Hence, the performance of these criteria in small and medium sized problems will be suboptimal.

Theorems 2 implies that  $\lim_{\sigma^2 \rightarrow 0} \Gamma_0 = \infty$  is a necessary condition for high SNR consistency. We next establish the necessity of  $\lim_{\sigma^2 \rightarrow 0} \sigma^2\Gamma_0 = 0$  for high SNR consistency.

**Theorem 3.** Consider a matrix  $\mathbf{X}$  which satisfies  $\text{spark}(\mathbf{X}) > 2k^*$ . Then for any  $k^*$ -sparse signal  $\beta$  with  $k^* \geq 1$ ,  $l_0$ -penalty is high SNR consistent only if  $\lim_{\sigma^2 \rightarrow 0} \sigma^2\Gamma_0 = 0$ .

*Proof.* Define  $\mathcal{J} = \mathcal{I}/i$ , where  $i \in \mathcal{I}$ . Since,  $\mathcal{J} \subset \mathcal{I}$  and  $\text{spark}(\mathbf{X}) > 2k^*$ , it follows that  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}})\mathbf{X}\beta \neq \mathbf{0}_n$ . Expanding  $L(\mathcal{J}) = \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}})\mathbf{y}\|_2^2 + \sigma^2\Gamma_0|\mathcal{J}|$  and applying Lemma 2, we have

$$PE \geq \mathbb{P}(L(\mathcal{J}) < L(\mathcal{I})) \geq \mathbb{P}(A < \Gamma_0\sigma^2), \quad (10)$$

where  $A = \mathbf{y}^T(\mathbf{P}_{\mathcal{I}} - \mathbf{P}_{\mathcal{J}})\mathbf{y} \sim \sigma^2\chi_1^2(\frac{\lambda}{\sigma^2})$  with  $\lambda = \|(\mathbf{P}_{\mathcal{I}} - \mathbf{P}_{\mathcal{J}})\mathbf{X}\beta\|_2^2 > 0$ . By Lemma 3,  $A \xrightarrow{P} \lambda$  as  $\sigma^2 \rightarrow 0$ . Suppose

that  $\lim_{\sigma^2 \rightarrow 0} \sigma^2\Gamma_0 = \lambda_1$ , where  $\lambda_1 > \lambda$ . Then,  $\exists \sigma_1^2 > 0$  such that  $\frac{\lambda_1 + \lambda}{2} < \sigma^2\Gamma_0 < \lambda_1$ ,  $\forall \sigma^2 < \sigma_1^2$ . This implies that

$$\begin{aligned} \mathbb{P}(A < \Gamma_0\sigma^2) &\geq \mathbb{P}(A < \frac{\lambda_1 + \lambda}{2}) = 1 - \mathbb{P}(A - \lambda > \frac{\lambda_1 - \lambda}{2}) \\ &\geq 1 - \mathbb{P}(|A - \lambda| > \frac{\lambda_1 - \lambda}{2}), \quad \forall \sigma^2 < \sigma_1^2. \end{aligned} \quad (11)$$

Since,  $A \xrightarrow{P} \lambda$  as  $\sigma^2 \rightarrow 0$ , for any  $\epsilon > 0$ ,  $\exists \sigma_2^2 > 0$  such that  $\mathbb{P}(|A - \lambda| > \frac{\lambda_1 - \lambda}{2}) \leq \epsilon$ ,  $\forall \sigma^2 < \sigma_2^2$ . Fix  $\sigma^2(\epsilon) = \min(\sigma_1^2, \sigma_2^2)$ . Then  $\forall \sigma^2 < \sigma^2(\epsilon)$ ,  $PE \geq 1 - \epsilon$ . Thus if  $\lambda_1 > \lambda$ , then  $\lim_{\sigma^2 \rightarrow 0} PE = 1$ . This implies that  $\lim_{\sigma^2 \rightarrow 0} \sigma^2\Gamma_0 < \lambda$  is a necessary condition for high SNR consistency. However, without *a priori* knowledge of non-zero entries of  $\beta$ ,  $\lambda$  is unknown. Hence,  $l_0$ -penalty is high SNR consistent only if  $\lim_{\sigma^2 \rightarrow 0} \sigma^2\Gamma_0 = 0$ . ■

*Remark 4.* The formulation of  $l_0$ -penalty given in (4) is exactly similar to that of MOS problems given in (1) except that the search space of MOS is a very small subset of the search space in  $l_0$ -penalty. This is reflected in the similarity of NSCs for MOS derived in [17] and Theorems 1-3 for subset selection. It is also true that different values of  $\Gamma_0$  gives EEF, NMDL etc. as special cases. Hence, Theorems 1-3 can be seen as an extension of the existing high SNR consistency results in [10], [17]–[19] to subset selection problems. However, the novelty of Theorems 1-3 lies in the fact that it explicitly takes into account the identifiability issues associated with subset selection in underdetermined linear models. These structural issues were not considered in [10], [17]–[19] which dealt with MOS in overdetermined linear regression models.

## IV. HIGH SNR CONSISTENCY OF CONVEX RELAXATION BASED SSPS

In this section, we derive NSCs on the tuning parameters  $\{\Gamma_i\}_{i=1}^3$  such that  $l_1$ -penalty,  $l_1$ -error and DS are high SNR consistent. Unlike the NP-hard  $l_0$ -penalty which is computationally infeasible except in small sized problems, the CR based SSPs discussed in this section and the greedy algorithms like OMP discussed in Section V can be implemented with polynomial complexity. Hence, these techniques are practically important. Unlike the high SNR consistency of  $l_0$ -penalty whose connections with the high SNR consistency in MOS problems we previously mentioned, the high SNR consistency of CR and greedy algorithms are not discussed in open literature to the best of our knowledge. We first discuss the  $l_1$ -penalty based SSP.

#### A. High SNR consistency of $l_1$ -penalty: Sufficient conditions

In this section, we discuss the high SNR behaviour of  $\hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \Gamma_1\sigma \|\mathbf{b}\|_1$  and  $\hat{\mathcal{I}} = \text{supp}(\hat{\beta})$ . This is a widely used SSP in high dimensional statistics.  $l_1$ -penalty is the convex program that is closest to the optimal but NP-hard  $l_0$ -penalty. Commonly used values of  $\Gamma_1$  include  $\Gamma_1 = 2\sqrt{2\log(p)}$  [34],  $\Gamma_1 = \sqrt{8(1 + \eta)\log(p - k^*)}$  [7],

$\Gamma_1 = 10\sqrt{\log(p)}$  [35] etc. Here,  $\eta > 0$  is a constant. The large sample consistency of  $l_1$ -penalty is also widely studied. For a fixed  $p$  and  $k^*$ , all values of  $\Gamma_1$  satisfying  $\frac{\Gamma_1}{n} \rightarrow 0$  and  $\frac{\Gamma_1}{n^{\frac{1+c}{2}}} \rightarrow \infty$  as  $n \rightarrow \infty$  results in large sample consistency under a set of regularity conditions [9].  $c$  depends on these regularity conditions. However, the consistency of  $l_1$ -penalty as  $\sigma^2 \rightarrow 0$  is not discussed in literature to the best of our knowledge. Next we state and prove the sufficient conditions for the high SNR consistency of  $l_1$ -penalty.

**Theorem 4.**  *$l_1$ -penalty is high SNR consistent for any matrix signal pair  $(\mathbf{X}, \beta)$  satisfying the ERC provided that the tuning parameter  $\Gamma_1$  satisfies  $\lim_{\sigma^2 \rightarrow 0} \Gamma_1 = \infty$  and  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_1 = 0$ .*

*Proof.* The proof of Theorem 4 is based on the following fundamental result proved in [Theorem 8, [4]].

**Lemma 4.** *Let  $\mathcal{J}$  be any index set satisfying ERC. If  $\mathbf{y}^{\mathcal{J}} = \mathbf{P}_{\mathcal{J}}\mathbf{y}$  satisfies  $\|\mathbf{X}^T(\mathbf{y} - \mathbf{y}^{\mathcal{J}})\|_{\infty} < \sigma \Gamma_1 (1 - \text{erc}(\mathbf{X}, \mathcal{J}))$ , then  $\hat{\beta}$  satisfies the following.*

A1).  $\text{supp}(\hat{\beta}) \subseteq \mathcal{J}$ .

A2).  $\hat{\beta}$  is the unique minimizer of  $l_1$ -penalty.

A3).  $\mathcal{T} = \{j : |\mathbf{b}^{\mathcal{J}}(j)| > \Gamma_1 \sigma \|(\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}})^{-1}\|_{\infty, \infty}\} \subseteq \text{supp}(\hat{\beta})$ , where  $\mathbf{b}^{\mathcal{J}} = \mathbf{X}_{\mathcal{J}}^{\dagger} \mathbf{y}$  is the LS estimate of  $\beta_{\mathcal{J}}$ .

In words, Lemma 4 states that if the correlation between the columns in  $\mathbf{X}$  and residual generated by the LS fit using the columns in  $\mathcal{J}$  is sufficiently low, then the support of solution to  $l_1$ -penalty will be contained in  $\mathcal{J}$ . Further,  $l_1$ -penalty does not miss indices that has sufficiently large values in the restricted LS estimate  $\mathbf{b}^{\mathcal{J}}$ . By the hypothesis of Theorem 4, the true support  $\mathcal{I}$  satisfies  $\text{erc}(\mathbf{X}, \mathcal{I}) < 1$ . Thus, if the event  $\mathcal{E}_1 = \{\|\mathbf{X}^T(\mathbf{y} - \mathbf{y}^{\mathcal{I}})\|_{\infty} < \Gamma_1 \sigma (1 - \text{erc}(\mathbf{X}, \mathcal{I}))\}$  is true, then  $\text{supp}(\hat{\beta}) \subseteq \mathcal{I}$ . That is,  $l_1$ -penalty does not make any false discoveries. If the event  $\mathcal{E}_2 = \{\forall j : |\mathbf{b}^{\mathcal{I}}(j)| > \Gamma_1 \sigma \|(\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}})^{-1}\|_{\infty, \infty}\} = \{|\mathcal{T}| = k^*\}$  is also true, then  $\text{supp}(\hat{\beta}) = \mathcal{I}$ . Thus  $\mathbb{P}(\hat{\mathcal{I}} = \mathcal{I}) \geq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)$ .

We first analyse the probability of the event  $\mathcal{E}_1$ . Note that  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{X}\beta = (\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{X}_{\mathcal{I}}\beta_{\mathcal{I}} = 0$ . Hence  $\|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{y}\|_{\infty} = \|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_{\infty}$ . Further,  $\|\mathbf{X}_j\|_2 = 1$  and Cauchy Schwartz inequality implies that  $\max_j |\mathbf{X}_j^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}| \leq \max_j \|\mathbf{X}_j\|_2 \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_2 = \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_2$ . Using these inequalities, we can bound  $\mathbb{P}(\mathcal{E}_1)$  as

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1) &= \mathbb{P}\left(\max_j |\mathbf{X}_j^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}| < \Gamma_1 \sigma (1 - \text{erc}(X, \mathcal{I}))\right) \\ &\geq \mathbb{P}(\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_2 < \Gamma_1 \sigma (1 - \text{erc}(X, \mathcal{I}))) \\ &= \mathbb{P}\left(\frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_2^2}{\sigma^2} < \Gamma_1^2 (1 - \text{erc}(X, \mathcal{I}))^2\right) \end{aligned} \quad (12)$$

Note that  $\frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_2^2}{\sigma^2} \sim \chi_{n-k^*}^2$  is a B.I.P R.V with distribution independent of  $\sigma^2$ . Hence, if the condition  $\lim_{\sigma^2 \rightarrow 0} \Gamma_1 = \infty$  in the hypotheses of Theorem 4 is satisfied, then the lower bound in (12) converges to 1. Hence,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) = 1$ .

Next, we analyse  $\mathbb{P}(\mathcal{E}_2)$ . Since  $\mathcal{I}$  is the correct support, it follows that  $\mathbf{b}^{\mathcal{I}} = \mathbf{X}_{\mathcal{I}}^{\dagger}(\mathbf{X}_{\mathcal{I}}\beta_{\mathcal{I}} + \mathbf{w}) = \beta_{\mathcal{I}} + \mathbf{X}_{\mathcal{I}}^{\dagger}\mathbf{w}$ . Since

$\mathbf{w} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$ , we have  $\mathbf{b}^{\mathcal{I}} \sim \mathcal{N}(\beta_{\mathcal{I}}, \sigma^2 (\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}})^{-1})$ . The set  $\mathcal{T}$  in A3) of Lemma 4 can be rewritten as  $\mathcal{T} = \{j : |\mathbf{b}^{\mathcal{I}}(j)| > \sigma c_j \Gamma_1 d_j\}$ , where  $c_j = \sqrt{((\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}})^{-1})_{j,j}}$  and  $d_j = \frac{\|(\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}})^{-1}\|_{\infty, \infty}}{c_j}$ . The NSC for the high SNR consistency of a threshold based SSP like this is given below.

**Lemma 5.** *Let  $\mathbf{z} \sim \mathcal{N}(\mathbf{u}, \sigma^2 \mathbf{C})$  and  $\mathcal{K} = \text{supp}(\mathbf{u})$ . Consider the threshold based estimator  $\hat{\mathcal{K}} = \{j : |\mathbf{z}_j| > \sigma \sqrt{\mathbf{C}_{j,j}} \Gamma\}$  of  $\mathcal{K}$ . Define the event false discovery  $\mathcal{F} = \{\exists j \in \hat{\mathcal{K}} \text{ and } j \notin \mathcal{K}\}$  and missed discovery  $\mathcal{M} = \{\exists j \notin \hat{\mathcal{K}} \text{ and } j \in \mathcal{K}\}$ . Then the following statements are true [8].*

L1).  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{F}) = 0$ , iff  $\lim_{\sigma^2 \rightarrow 0} \Gamma = \infty$ .

L2).  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{M}) = 0$ , iff  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma < \min_{j \in \mathcal{K}} \frac{|\mathbf{u}_j|}{\sqrt{\mathbf{C}_{j,j}}}$ .

Hence, if  $\Gamma_1$  satisfies  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_1 = 0$ , then by L2) of Lemma 5, all entries in  $\hat{\mathcal{I}}$  will be included in  $\mathcal{I}$  at high SNR. Mathematically,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2) = \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(|\mathcal{T}| = k^*) = 1$ . Since,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) = 1$  and  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2) = 1$ , it follows that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\hat{\mathcal{I}} = \mathcal{I}) \geq \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = 1$ . ■

*B. On the choice of SNR adaptation in  $\Gamma_1$ .*

Theorem 4 states that all SNR adaptations on  $\Gamma_1$  satisfying  $\lim_{\sigma^2 \rightarrow 0} \Gamma_1 = \infty$  and  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_1 = 0$  results in the high SNR consistency of  $l_1$ -penalty. However, the choice of SNR adaptation has profound influence on the performance of  $l_1$ -penalty in the moderate to high SNR range. In this section, we derive convergence rates for  $\mathbb{P}(\mathcal{E}_1)$  and  $\mathbb{P}(\mathcal{E}_2)$  discussed in the proof of Theorem 4. First consider the event  $\mathcal{E}_1 = \{\|\mathbf{X}^T(\mathbf{y} - \mathbf{y}^{\mathcal{I}})\|_{\infty} < \Gamma_1 \sigma (1 - \text{erc}(\mathbf{X}, \mathcal{I}))\}$ . Following (12), we have

$$\mathbb{P}(\mathcal{E}_1) \geq 1 - \mathbb{P}\left(A > \Gamma_1^2 (1 - \text{erc}(X, \mathcal{I}))^2\right), \quad (13)$$

where  $A \sim \chi_{n-k^*}^2$ . Let  $X \sim \chi_k^2$  and  $a^2 > k$ . Then by Lemma 10 in [17], we have

$$\mathbb{P}(X > a^2) \leq \frac{\exp(\frac{k}{2})}{k^{\frac{k}{2}}} \exp\left(\frac{-1}{2}[a^2 - k \log(a^2)]\right). \quad (14)$$

Let  $b_1 = 1 - \text{erc}(\mathbf{X}, \mathcal{I})$  and  $b_2 = \frac{\exp(\frac{n-k^*}{2})}{(n-k^*)^{\frac{n-k^*}{2}}}$ . Applying (14) in (13) gives,

$$\mathbb{P}(\mathcal{E}_1) \geq 1 - b_2 \exp\left(\frac{-1}{2}[\Gamma_1^2 b_1^2 - (n - k^*) \log(\Gamma_1^2 b_1^2)]\right). \quad (15)$$

The R.H.S of inequality in (15) is independent of  $\sigma^2$  for the SNR independent  $\Gamma_1$  discussed in literature. Further, the inequality (15) converges to one faster as the growth of  $\Gamma_1$  increases. Let  $\Gamma_1 = \frac{1}{\sigma^\alpha}$  be the SNR adaptation in  $\Gamma_1$ . This adaptation satisfies Theorem 4 if  $0 < \alpha < 1$ . The convergence rate of  $\mathbb{P}(\mathcal{E}_1)$  will be faster for  $\alpha_1$  than that of  $\alpha_2$  if  $\alpha_1 > \alpha_2$ .

Next consider the event  $\mathcal{E}_2 = \{\forall j : |\mathbf{b}^{\mathcal{I}}(j)| > \Gamma_1 \sigma c_j d_j\}$ , where  $\mathbf{b}^{\mathcal{I}} = \mathbf{X}_{\mathcal{I}}^{\dagger} \mathbf{y}$ ,  $c_j = \sqrt{((\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}})^{-1})_{j,j}}$  and  $d_j = \frac{\|(\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}})^{-1}\|_{\infty, \infty}}{c_j}$  as used in the proof of Theorem 4. The



following set of inequalities follows directly from union bound and the  $\mathbf{b}^{\mathcal{I}}(j) \sim \mathcal{N}(\beta_j, \sigma^2 c_j^2)$  distribution of  $\mathbf{b}^{\mathcal{I}}(j)$ .

$$\begin{aligned} \mathbb{P}(\mathcal{E}_2) &= \mathbb{P}\left(\bigcap_{j \in [k^*]} |\mathbf{b}^{\mathcal{I}}(j)| > \sigma c_j \Gamma_1 d_j\right) \\ &\geq 1 - \sum_{j=1}^{k^*} \mathbb{P}(|\mathbf{b}^{\mathcal{I}}(j)| < \sigma c_j \Gamma_1 d_j) \\ &= 1 - \sum_{j=1}^{k^*} \left[ Q(-\Gamma_1 d_j - \frac{\beta_j}{\sigma c_j}) - Q(\Gamma_1 d_j - \frac{\beta_j}{\sigma c_j}) \right]. \end{aligned} \quad (16)$$

Applying  $Q(x) > 0, \forall x$ , gives

$$\mathbb{P}(\mathcal{E}_2) \geq 1 - \sum_{j=1}^{k^*} Q(-\Gamma_1 d_j - \frac{\beta_j}{\sigma c_j}). \quad (17)$$

For the ease of exposition assume that  $\beta_j < 0, \forall j \in \mathcal{I}$ . Since,

$\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_1 = 0$ , we have  $\lim_{\sigma^2 \rightarrow 0} (-\Gamma_1 d_j - \frac{\beta_j}{\sigma c_j}) = \infty$ . Hence,

$\exists \sigma_1^2 > 0$  such that  $-\Gamma_1 d_j - \frac{\beta_j}{\sigma c_j} > 2, \forall j$ . Using the bound

$Q(x) \leq \frac{1}{2} \exp\left(-\frac{x^2}{2}\right), \forall x > 2$ , we have

$$\mathbb{P}(\mathcal{E}_2) \geq 1 - \frac{1}{2} \sum_{j=1}^{k^*} \exp\left(-\frac{\left(-\Gamma_1 d_j - \frac{\beta_j}{\sigma c_j}\right)^2}{2}\right), \quad (18)$$

$\forall \sigma^2 < \sigma_1^2$ . Unlike the bound (15) on  $\mathbb{P}(\mathcal{E}_1)$ , the R.H.S in (18) increases with the signal strength  $|\beta_j|$ . Further, the convergence rate of  $\mathbb{P}(\mathcal{E}_2)$  decreases with the increase in the rate at which  $\Gamma_1$  increase to  $\infty$ . For  $\Gamma_1 = \frac{1}{\sigma^\alpha}$ , the convergence rate of  $\mathbb{P}(\mathcal{E}_2)$  decreases with increase in  $\alpha$ .

We now make the following observations on the choice of SNR adaptations based on (15) and (18). Consider SNR adaptations of the form  $\Gamma_1 = \frac{1}{\sigma^\alpha}$ . When signal strength is low, i.e.,  $\beta_j$  is low for some  $j \in \mathcal{I}$ , it is reasonable to choose slow rates for  $\Gamma_1$  like  $\alpha = 0.1$ . This will ensure the increase of  $\mathbb{P}(\mathcal{E}_1)$  to one at a descent rate without causing significant decrease in the convergence rates of  $\mathbb{P}(\mathcal{E}_2)$ . However, when the signal strength is high, i.e.,  $\beta_j$  is high for all  $j \in \mathcal{I}$ ,  $\mathbb{P}(\mathcal{E}_2)$  will be close to one for moderate values of SNR for most values of  $0 < \alpha < 1$ . Then the gain in the convergence rate of  $\mathbb{P}(\mathcal{E}_1)$  by allowing a larger value of  $\alpha$  will overpower the slight decrease in the convergence rate in  $\mathbb{P}(\mathcal{E}_2)$ . Hence, when signal strength is high, one can choose faster SNR adaptations like  $\alpha = 0.5$ .

### C. High SNR consistency of $l_1$ -penalty: Necessary conditions

In the following, we establish the necessity of SNR adaptations detailed in Theorem 4 for high SNR consistency.

**Theorem 5.** *Suppose  $\exists \mathcal{J} \supset \mathcal{I}$  such that the matrix support pair  $(\mathbf{X}, \mathcal{J})$  satisfy ERC. Then  $l_1$ -penalty is high SNR consistent only if  $\lim_{\sigma^2 \rightarrow 0} \Gamma_1 = \infty$ .*

*Proof.* Let  $\mathcal{J} \supset \mathcal{I}$  be an index set satisfying ERC. Define the events  $\mathcal{E}_1 : \{\|\mathbf{X}^T(\mathbf{y} - \mathbf{y}^{\mathcal{J}})\|_\infty < \Gamma_1 \sigma (1 - \text{erc}(\mathbf{X}, \mathcal{J}))\}$  and

$\mathcal{E}_2 : \{|\mathcal{T}| = |\mathcal{J}|\}$ , where  $\mathcal{T} = \{j : |\mathbf{b}^{\mathcal{J}}(j)| > c_j \sigma \Gamma_1 d_j\}$ ,  $c_j = \sqrt{((\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}})^{-1})_{j,j}}$  and  $d_j = \frac{\|(\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}})^{-1}\|_{\infty, \infty}}{c_j}$ .

$\mathbf{y}^{\mathcal{J}} = \mathbf{P}_{\mathcal{J}} \mathbf{y}$  and  $\mathbf{b}^{\mathcal{J}} = \mathbf{X}_{\mathcal{J}}^\dagger \mathbf{y}$  are the same as in Lemma 4. If both these events are true, then by Lemma 4,  $\hat{\mathcal{I}} = \text{supp}(\hat{\beta}) = \mathcal{J} \supset \mathcal{I}$ . Hence,  $PE = \mathbb{P}(\hat{\mathcal{I}} \neq \mathcal{I}) \geq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)$ . Since  $\mathcal{I} \subset \mathcal{J}$ ,  $\mathbf{y} - \mathbf{y}^{\mathcal{J}} = (\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{y} = (\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{w}$ . Replacing  $\mathcal{I}$  with  $\mathcal{J}$  in (12), we have

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1) &= \mathbb{P}\left(\max_j |\mathbf{X}_{\mathcal{J}}^T (\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{w}| < \sigma \Gamma_1 (1 - \text{erc}(\mathbf{X}, \mathcal{J}))\right) \\ &\geq \mathbb{P}\left(A < \Gamma_1^2 (1 - \text{erc}(\mathbf{X}, \mathcal{J}))^2\right), \end{aligned} \quad (19)$$

where  $A = \frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{w}\|_2^2}{\sigma^2} \sim \chi_{n-|\mathcal{J}|}^2$  is a R.V with distribution independent of  $\sigma^2$  and support in  $(0, \infty)$ . Hence, as long as  $\lim_{\sigma^2 \rightarrow 0} \Gamma_1 > 0$ ,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}\left(A < \Gamma_1^2 (1 - \text{erc}(\mathbf{X}, \mathcal{J}))^2\right) > 0$ , which in turn imply that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) > 0$ .

We next consider  $\mathbb{P}(\mathcal{E}_2)$ . Since  $\mathcal{I} \subset \mathcal{J}$ , we have  $\mathbf{X}_{\mathcal{I}} \beta_{\mathcal{I}} = \mathbf{X}_{\mathcal{J}} \beta_{\mathcal{J}}$  with appropriate zero entries in  $\beta_{\mathcal{J}}$ . Thus,  $\mathbf{b}^{\mathcal{J}} \sim \mathcal{N}(\beta_{\mathcal{J}}, \sigma^2 (\mathbf{X}_{\mathcal{J}}^T \mathbf{X}_{\mathcal{J}})^{-1})$ . Hence,  $|\mathcal{T}| = |\mathcal{J}|$  iff a false discovery is made in the thresholding procedure which gives  $\mathcal{T}$ . From Lemma 5, it follows that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2) = \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(|\mathcal{T}| = |\mathcal{J}|) > 0$  as long as  $\lim_{\sigma^2 \rightarrow 0} \Gamma_1 < \infty$ .

A careful analysis of the events  $\mathcal{E}_1$  and  $\mathcal{E}_2$  reveals that  $\mathcal{E}_1$  depends only on the component  $(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}) \mathbf{w}$  of  $\mathbf{w}$  and  $\mathcal{E}_2$  depends only on the component  $\mathbf{P}_{\mathcal{J}} \mathbf{w}$ . Since, these two components are orthogonal and  $\mathbf{w}$  is Gaussian, it follows that  $\mathcal{E}_1$  and  $\mathcal{E}_2$  are mutually independent, i.e.,  $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \mathbb{P}(\mathcal{E}_1) \mathbb{P}(\mathcal{E}_2)$ . Since,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) > 0$  and  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2) > 0$ , it follows that  $\lim_{\sigma^2 \rightarrow 0} PE \geq \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2) > 0$ , unless  $\lim_{\sigma^2 \rightarrow 0} \Gamma_1 = \infty$ . ■

It must be mentioned that an index set  $\mathcal{J} \supset \mathcal{I}$  satisfying ERC need not exist in all situations where  $\mathcal{I}$  satisfy ERC. In that sense, Theorem 5 is less general than Theorem 4. Nevertheless, Theorem 5 is applicable in many practical settings. For example, if  $\mathbf{X} \in \mathbb{R}^{n \times p}$  is orthonormal, then  $\mathbf{X}$  satisfies ERC for all possible index sets  $\mathcal{J} \subseteq [p]$ . Similarly, if  $\mathbf{X}$  satisfies the MIC of order  $j$ , i.e.,  $\mu_{\mathbf{X}} \leq \frac{1}{2j-1}$  and  $j > k^*$ , then  $\mathbf{X}$  satisfies ERC for all  $j > k^*$  sized index sets. In both these situations, an index set (in fact many)  $\mathcal{J} \supset \mathcal{I}$  satisfying ERC exists and  $l_1$ -penalty will be inconsistent without the required SNR adaptation. Theorem 5 proves that the values of  $\Gamma_1$  discussed in literature makes  $l_1$ -penalty high SNR inconsistent even in the simple case of orthonormal design matrix. Next we establish the necessity of  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_1 = 0$  for high SNR consistency.

**Theorem 6.** *Suppose that the matrix support pair  $(\mathbf{X}, \mathcal{I})$  satisfy ERC and  $k^* \geq 1$ . Then,  $l_1$ -penalty will be high SNR consistent only if  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_1 = 0$ .*

*Proof.* Let  $\mathcal{J}$  be any index set satisfying  $\mathcal{J} \subset \mathcal{I}$ . Since,  $\mathcal{I}$  satisfy ERC,  $\mathcal{J}$  will also satisfy ERC. Consider the event  $\mathcal{E} : \{\|\mathbf{X}^T(\mathbf{y} - \mathbf{y}^{\mathcal{J}})\|_\infty < \Gamma_1 \sigma (1 - \text{erc}(\mathbf{X}, \mathcal{J}))\}$ , where  $\mathbf{y}^{\mathcal{J}} =$

$\mathbf{P}_{\mathcal{J}}\mathbf{y}$ . If  $\mathcal{E}$  is true, then by Lemma 4,  $\hat{\mathcal{I}} = \text{supp}(\hat{\beta}) \subseteq \mathcal{J} \subset \mathcal{I}$ . Thus,  $PE \geq \mathbb{P}(\mathcal{E})$ . The following bound on  $\mathbb{P}(\mathcal{E})$  follows from Cauchy Schwartz inequality and the unit  $l_2$  norm of  $\mathbf{X}_j$ .

$$\begin{aligned} \mathbb{P}(\mathcal{E}) &= \mathbb{P}\left(\max_j |\mathbf{X}_j^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}})\mathbf{y}| < \sigma\Gamma_1(1 - \text{erc}(\mathbf{X}, \mathcal{J}))\right) \\ &\geq P(\|\mathbf{I}_n - \mathbf{P}_{\mathcal{J}}\mathbf{y}\|_2 < \sigma\Gamma_1(1 - \text{erc}(\mathbf{X}, \mathcal{J}))) \\ &= \mathbb{P}(\sigma^2 A < \sigma^2\Gamma_1^2(1 - \text{erc}(\mathbf{X}, \mathcal{J}))^2), \end{aligned} \quad (20)$$

where  $A = \frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}})\mathbf{y}\|_2^2}{\sigma^2} \sim \chi_{n-|\mathcal{J}|}^2\left(\frac{\lambda}{\sigma^2}\right)$  and  $\lambda = \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}})\mathbf{X}\beta\|_2^2 > 0$ . By Lemma 3,  $\sigma^2 A \xrightarrow{P} \lambda$  as  $\sigma^2 \rightarrow 0$ . Hence, if  $\lim_{\sigma^2 \rightarrow 0} \sigma^2\Gamma_1^2(1 - \text{erc}(\mathbf{X}, \mathcal{J}))^2 > \lambda$ , then as shown in the proof of Theorem 3,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\sigma^2 A < \sigma^2\Gamma_1^2(1 - \text{erc}(\mathbf{X}, \mathcal{J}))^2) = 1$ . However,  $\lambda$  is unknown. Thus, to satisfy  $\lim_{\sigma^2 \rightarrow 0} \sigma^2\Gamma_1^2(1 - \text{erc}(\mathbf{X}, \mathcal{J}))^2 < \lambda$ , it is necessary that  $\lim_{\sigma^2 \rightarrow 0} \sigma^2\Gamma_1^2 = 0$  which is equivalent to  $\lim_{\sigma^2 \rightarrow 0} \sigma\Gamma_1 = 0$ . ■

A widely used formulation of  $l_1$ -penalty is given by  $\hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1$  which is equivalent to the formulation in this article by setting  $\lambda = \Gamma_1\sigma$ . In this formulation,  $l_1$ -penalty is high SNR consistent if  $\lim_{\sigma^2 \rightarrow 0} \frac{\lambda}{\sigma} = \infty$  and  $\lim_{\sigma^2 \rightarrow 0} \lambda = 0$ . An interesting case is that of a fixed  $\sigma$  independent  $\lambda$  like  $\lambda = 0.1$ . This choice of  $\lambda$  satisfy  $\lim_{\sigma^2 \rightarrow 0} \frac{\lambda}{\sigma} = \infty$  which is a necessary condition for high SNR consistency. However, the satisfiability of the necessary condition in Theorem 6 depends upon on the signal  $\beta$  (Please see the proof of Theorem 6). Hence, when *a priori* knowledge of  $\beta$  is not available, a fixed regularization parameter is not advisable from the vantage point of high SNR consistency.

#### D. High SNR consistency of $l_1$ -error

We next discuss the high SNR behaviour of  $\hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{b}\|_1$ , subject to  $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2 \leq \Gamma_2\sigma$  and  $\hat{\mathcal{I}} = \text{supp}(\hat{\beta})$ .  $l_1$ -penalty is the Lagrangian of the constrained optimization problem given by  $l_1$ -error. The performance of  $l_1$ -error is dictated by the choice of  $\Gamma_2$ . Commonly used choice of  $\Gamma_2$  include  $\Gamma_2 = \sqrt{n + 2\sqrt{2n}}$  [36],  $\Gamma_2 = \sqrt{n + 2\sqrt{n \log(n)}}$  [37] etc. A high SNR analysis of  $l_1$ -error in terms of variable selection properties is not available in open literature to the best of our knowledge. The following theorem states the sufficient conditions for  $l_1$ -error to be high SNR consistent.

**Theorem 7.**  $l_1$ -error is high SNR consistent for any matrix support pair  $(\mathbf{X}, \mathcal{I})$  satisfying the ERC provided that the tuning parameter  $\Gamma_2$  satisfies  $\lim_{\sigma^2 \rightarrow 0} \Gamma_2 = \infty$  and  $\lim_{\sigma^2 \rightarrow 0} \sigma\Gamma_2 = 0$ .

*Proof.* The proof of Theorem 7 is based on the result in [Theorem 14, [4]] regarding the minimizers of  $l_1$ -error.

**Lemma 6.** Let  $\mathcal{J}$  be any index set satisfying ERC. If  $\mathbf{y}^{\mathcal{J}} = \mathbf{P}_{\mathcal{J}}\mathbf{y}$  satisfies

$$\Gamma_2^2\sigma^2 \geq \|\mathbf{y} - \mathbf{y}^{\mathcal{J}}\|_2^2 + \frac{\|\mathbf{X}^T(\mathbf{y} - \mathbf{y}^{\mathcal{J}})\|_\infty^2 \|\mathbf{X}_{\mathcal{J}}^\dagger\|_{2,1}^2}{(1 - \text{erc}(\mathbf{X}, \mathcal{J}))^2}, \quad (21)$$

then  $\hat{\beta}$  satisfies the following.

A1).  $\text{supp}(\hat{\beta}) \subseteq \mathcal{J}$ .

A2).  $\hat{\beta}$  is the unique minimizer of  $l_1$ -error.

A3).  $\mathcal{T} = \{j : |\mathbf{b}^{\mathcal{J}}(j)| > \Gamma_2\sigma\|\mathbf{X}_{\mathcal{J}}^\dagger\|_{2,2}\} \subseteq \text{supp}(\hat{\beta})$ .

Here  $\mathbf{b}^{\mathcal{J}} = \mathbf{X}_{\mathcal{J}}^\dagger\mathbf{y}$  is the LS estimate of  $\beta_{\mathcal{J}}$ .

In words, Lemma 6 states that if the residual between  $\mathbf{y}$  and the LS fit of  $\mathbf{y}$  using the columns in  $\mathbf{X}_{\mathcal{J}}$  has sufficiently low correlation with the columns in  $\mathbf{X}$  and sufficiently low  $l_2$  norm, then the support of the solution to  $l_1$ -error will be contained in  $\mathcal{J}$ . Further,  $l_1$ -error does not miss indices that have sufficiently large values in the restricted LS estimate  $\mathbf{b}^{\mathcal{J}}$ . By the hypothesis of Theorem 7, the true support  $\mathcal{I}$  satisfies  $\text{erc}(\mathbf{X}, \mathcal{I}) < 1$ . Thus, if the event  $\mathcal{E}_1 = \{(21) \text{ is satisfied}\}$ , then  $\hat{\mathcal{I}} = \text{supp}(\hat{\beta}) \subseteq \mathcal{I}$ . If the event  $\mathcal{E}_2 = \{\forall j : |\mathbf{b}^{\mathcal{I}}(j)| > \Gamma_2\sigma\|\mathbf{X}_{\mathcal{I}}^\dagger\|_{2,2}\} = \{|\mathcal{T}| = k^*\}$  is also true, then  $\text{supp}(\hat{\beta}) = \mathcal{I}$ . Thus  $\mathbb{P}(\hat{\mathcal{I}} = \mathcal{I}) \geq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)$ .

We first analyse the probability of the event  $\mathcal{E}_1$ . By Cauchy Schwartz inequality and the fact that  $\|\mathbf{X}_j\|_2 = 1$ , we have  $|\mathbf{X}_j^T(\mathbf{y} - \mathbf{y}^{\mathcal{J}})| \leq \|\mathbf{y} - \mathbf{y}^{\mathcal{J}}\|_2, \forall j$ . Thus,  $\|\mathbf{X}^T(\mathbf{y} - \mathbf{y}^{\mathcal{J}})\|_\infty^2 \leq \|\mathbf{y} - \mathbf{y}^{\mathcal{J}}\|_2^2$ . Hence,  $\Gamma_2^2\sigma^2 > \|\mathbf{y} - \mathbf{y}^{\mathcal{J}}\|_2^2 a_{\mathcal{I}}$ , where  $a_{\mathcal{I}} = \left(1 + \frac{\|\mathbf{X}_{\mathcal{I}}^\dagger\|_{2,1}^2}{(1 - \text{erc}(\mathbf{X}, \mathcal{I}))^2}\right)$  implies (21). Thus,

$$\mathbb{P}(\mathcal{E}_1) \geq \mathbb{P}\left(\frac{\|\mathbf{y} - \mathbf{y}^{\mathcal{I}}\|_2^2}{\sigma^2} < \Gamma_2^2 a_{\mathcal{I}}^{-1}\right) \quad (22)$$

Note that  $\mathbf{y} - \mathbf{y}^{\mathcal{I}} = (\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{y} = (\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}$ . Hence,  $A = \frac{\|\mathbf{y} - \mathbf{y}^{\mathcal{I}}\|_2^2}{\sigma^2} \sim \chi_{n-k^*}^2$ . Since,  $\chi_{n-k^*}^2$  is a B.I.P R.V with  $\sigma^2$  independent distribution, it follows that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(A < \Gamma_2^2 a_{\mathcal{I}}^{-1}) = 1$  if  $\lim_{\sigma^2 \rightarrow 0} \Gamma_2 = \infty$ . This implies that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) = 1$ .

Next we consider the event  $\mathcal{E}_2$ . The index set  $\mathcal{T}$  in Lemma 6 can be rewritten as  $\mathcal{T} = \{j : |\mathbf{b}^{\mathcal{I}}(j)| > \sigma c_j \Gamma_1 d_j\}$ , where  $c_j = \sqrt{(\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}})^{-1}}$  and  $d_j = \frac{\|\mathbf{X}_{\mathcal{J}}^\dagger\|_{2,2}}{c_j}$ . The event  $\{|\mathcal{T}| = k^*\}$  happens iff there is no missed discovery in the thresholding procedure generating  $\mathcal{T}$ . Then it follows from  $\lim_{\sigma^2 \rightarrow 0} \sigma\Gamma_1 = 0$  and L2) of Lemma 5 that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2) = \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(|\mathcal{T}| = k^*) = 1$ . Since,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) = 1$  and  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2) = 1$ , it follows that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\hat{\mathcal{I}} = \mathcal{I}) \geq \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = 1$ . ■

The following theorem states that the SNR adaptations outlined in Theorem 7 are necessary for high SNR consistency.

**Theorem 8.** The following statements regarding the high SNR consistency of  $l_1$ -error are true.

1). Suppose  $\exists \mathcal{J} \supset \mathcal{I}$  such that the matrix support pair  $(\mathbf{X}, \mathcal{J})$  satisfy ERC. Then  $l_1$ -error is high SNR consistent only if

$$\lim_{\sigma^2 \rightarrow 0} \Gamma_2 = \infty.$$

2). Suppose that the matrix support pair  $(\mathbf{X}, \mathcal{I})$  satisfy ERC



and  $k^* \geq 1$ . Then,  $l_1$ -error will be high SNR consistent only if  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_2 = 0$ .

*Proof.* Similar to Theorem 5 and Theorem 6. ■

Note that the values of  $\Gamma_2$  discussed in literature do not satisfy the NSCs outlined in Theorems 7 and 8. Hence,  $l_1$ -error with these values of  $\Gamma_2$  will be inconsistent at high SNR.

### E. Analysis of Dantzig selector based SSP

Here, we discuss the high SNR behaviour of  $\hat{\beta}$  given by  $\hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{b}\|_1$ , subject to  $\|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{b})\|_\infty \leq \Gamma_3\sigma$ . and  $\hat{\mathcal{I}} = \text{supp}(\hat{\beta})$ . The properties of DS is determined largely by the hyper parameter  $\Gamma_3$ . Commonly used values include  $\Gamma_3 = \sqrt{2 \log(p)}$  [2],  $\Gamma_3 = (\frac{3}{2} + \sqrt{2 \log(p)})$  [38] etc. No high SNR consistency results for DS is reported in open literature to the best of our knowledge. Next we state and prove the NSCs for the high SNR consistency of DS when  $\mathbf{X}$  is orthonormal.

**Theorem 9.** For an orthonormal design matrix  $\mathbf{X}$ , DS is high SNR consistent iff  $\lim_{\sigma^2 \rightarrow 0} \Gamma_3 = \infty$  and  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_3 < \min_{j \in \mathcal{I}} |\beta_j|$ .

*Proof.* When  $\mathbf{X}$  is orthonormal, the solution to DS is given by  $\hat{\beta}_j = (|(\mathbf{X}^T \mathbf{y})_j| - \Gamma_3 \sigma)_+ \text{sign}((\mathbf{X}^T \mathbf{y})_j)$ ,  $\forall j$ . Note that  $(x)_+ = x$  if  $x > 0$  and  $(x)_+ = 0$  if  $x \leq 0$ . Thus  $\hat{\mathcal{I}}$  is obtained by thresholding the vector  $|\mathbf{X}^T \mathbf{y}|$  at level  $\sigma \Gamma_3$ . Since,  $\mathbf{X}$  is orthonormal, we have  $\mathbf{X}^T \mathbf{y} \sim \mathcal{N}(\beta, \sigma^2 \mathbf{I}_p)$ . The proof now follows directly from Lemma 5. ■

Note that no *a priori* knowledge of  $\beta$  is available. Hence, to achieve consistency, it is necessary that  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_3 = 0$ . It follows directly from Theorem 9 that the values of  $\Gamma_3$  discussed in literature are inconsistent for orthonormal matrices. This implies the inconsistency of these tuning parameters in regression classes based on  $\mu_{\mathbf{X}}$  and ERC which includes orthonormal matrices too. We now make an observation regarding the NSCs developed for  $l_0$ -penalty,  $l_1$ -error,  $l_1$ -penalty and DS.

*Remark 5.* The SNR adaptations prescribed for high SNR consistency have many similarities. Even though  $l_0$ -penalty requires  $\Gamma_0 \sigma^2 \rightarrow 0$  whereas other algorithms requires  $\Gamma_i \sigma \rightarrow 0$ , the effective regularization parameter, i.e.,  $\lambda_0 = \Gamma_0 \sigma^2$  for  $l_0$ -penalty and  $\lambda_i = \Gamma_i \sigma$  for other algorithms satisfies  $\lambda_i \rightarrow 0$  as  $\sigma^2 \rightarrow 0$ . In the absence of noise (i.e.  $\sigma^2 = 0$ ) equality constrained optimization problems (2) and (3) will correctly recover the support of  $\beta$  under spark and ERC assumptions respectively. Further, when the effective regularization parameter  $\lambda_i \rightarrow 0$ ,  $l_0$ -penalty automatically reduces to (2), whereas,  $l_1$ -penalty,  $l_1$ -error and DS reduces to (3). Hence, the condition  $\lambda_i \rightarrow 0$  as  $\sigma^2 \rightarrow 0$  is a natural choice to transition from the formulations for noisy data to the equality constrained  $l_0$  or  $l_1$  minimization ideal for noiseless data.

## V. ANALYSIS OF ORTHOGONAL MATCHING PURSUIT

OMP [21]–[23] is one of most popular techniques in the class of greedy algorithms to solve CS problems. Unlike the CR techniques like  $l_1$ -penalty which has a computational complexity  $O(np^2)$ , OMP has a complexity of only  $O(npk^*)$ .

Consequently, OMP is more easily scalable to large scale problems than CR techniques. Further, the performance guarantees for OMP are only slightly weaker compared to CR techniques. An algorithmic description of OMP is given below.

- Step 1: Initialize the residual  $\mathbf{r}^0 = \mathbf{y}$ . Support estimate  $\mathcal{J}^0 = \emptyset$ . Iteration counter  $i = 1$ ;
- Step 2: Find the column index most correlated with the current residual  $\mathbf{r}^{i-1}$ , i.e.,  $t_i = \arg \max_{t \in [p]} |\mathbf{X}_t^T \mathbf{r}^{i-1}|$ .
- Step 3: Update support estimate:  $\mathcal{J}^i = \mathcal{J}^{i-1} \cup t_i$ .
- Step 4: Update residual:  $\mathbf{r}^i = (\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^i}) \mathbf{y}$ .
- Step 5: Repeat Steps 2-4, if stopping condition (SC) is not met, else, output  $\hat{\mathcal{I}} = \mathcal{J}^i$ .

The properties of OMP is determined by the SC. A large body of literature regarding OMP assumes *a priori* knowledge of sparsity level of  $\beta$ , i.e.,  $k^*$  and run  $k^*$  iterations of OMP [21], [22]. When  $k^*$  is not known, two popular SCs for OMP are discussed in literature. One SC called residual power based stopping condition (RPSC) terminate iterations when the residual power becomes too low (i.e.,  $\|\mathbf{r}^i\|_2 < \sigma \Gamma_4$ ) and other SC called residual correlation based stopping condition (RCSC) terminate iterations when the maximum correlation of columns in  $\mathbf{X}$  with the residual becomes too low (i.e.,  $\|\mathbf{X}^T \mathbf{r}^i\|_\infty < \sigma \Gamma_5$ ). A commonly used value of  $\Gamma_4$  is  $\Gamma_4 = \sqrt{n + 2\sqrt{n \log(n)}}$  and that of  $\Gamma_5$  is  $\Gamma_5 = \sqrt{2(1 + \eta) \log(p)}$  [23]. Here  $\eta > 0$  is a constant. The following theorems state the sufficient conditions for OMP with RPSC and RCSC to be high SNR consistent.

**Theorem 10.** OMP with RPSC is high SNR consistent for any matrix  $\mathbf{X}$  and signal  $\beta$  satisfying the ERC provided that the hyper parameter  $\Gamma_4$  satisfies  $\lim_{\sigma^2 \rightarrow 0} \Gamma_4 = \infty$  and  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_4 = 0$ .

**Theorem 11.** OMP with RCSC is high SNR consistent for any matrix  $\mathbf{X}$  and signal  $\beta$  satisfying the ERC provided that the hyper parameter  $\Gamma_5$  satisfies  $\lim_{\sigma^2 \rightarrow 0} \Gamma_5 = \infty$  and  $\lim_{\sigma^2 \rightarrow 0} \sigma \Gamma_5 = 0$ .

### A. Proofs of Theorem 10 and Theorem 11

Let us consider the two processes- OMP iterating without SC (P1) and verification of the SC (P2) separately. Specifically P1 returns a set of indexes in order, say  $\{t_1, t_2, \dots\}$  and P2 returns a single index  $j$  indicating where to stop. Then, the support estimate is given by  $\hat{\mathcal{I}} = \{t_1, \dots, t_j\}$  and the indices after  $j$ , i.e.,  $\{t_{j+1}, \dots\}$  will be discarded. Let  $\mathcal{E}_1$  denotes the event  $\{t_1, \dots, t_{k^*}\} = \mathcal{I}$ , i.e., the first  $k^*$  iterations of OMP returns all the  $k^*$  indices in  $\mathcal{I}$  and  $\mathcal{E}_2$  denotes the event  $\{P2 \text{ returns } k^*\}$ . Then  $\mathbb{P}(\hat{\mathcal{I}} = \mathcal{I}) = \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)$ .

Let  $N_i = \|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}}) \mathbf{w}\|_\infty$  denotes the maximum correlation between the columns in  $\mathbf{X}$  and noise component in the current residual  $\mathbf{r}^{i-1}$  and  $\beta_{\min} = \min_{j \in \mathcal{I}} |\beta_j|$  denotes the minimum non-zero value in  $\beta$ . Then, using the analysis in Section V of [23],  $N_i < c_{\mathcal{I}} \beta_{\min}$ , where  $c_{\mathcal{I}} = \frac{(1 - \text{erc}(\mathbf{X}, \mathcal{I})) \lambda_{\min}(\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}})}{2\sqrt{k^*}}$  is a sufficient condition for selecting an index from  $\mathcal{I}$  in the  $i^{\text{th}}$  iteration ( $\forall i \leq k^*$ ). Since,  $\|\mathbf{X}_j\|_2 = 1$ , it follows that  $N_i \leq \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}}) \mathbf{w}\|_2$ . Thus,  $\mathbb{P}(\mathcal{E}_1) \geq \mathbb{P}(\bigcap_{i=1, \dots, k^*} \{ \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}}) \mathbf{w}\|_2 < c_{\mathcal{I}} \beta_{\min} \})$ .

One can bound  $\mathbb{P}(\mathcal{E}_1^C)$  using union bound and the inequality  $\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}})\mathbf{w}\|_2 \leq \|\mathbf{w}\|_2$  as

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1^C) &\leq \mathbb{P}\left(\bigcup_{i=1, \dots, k^*} \{ \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}})\mathbf{w}\|_2 > c_{\mathcal{I}}\beta_{\min} \}\right) \\ &\leq \sum_{i=1}^{k^*} \mathbb{P}\left(\frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}})\mathbf{w}\|_2^2}{\sigma^2} > \frac{c_{\mathcal{I}}^2\beta_{\min}^2}{\sigma^2}\right) \\ &\leq \sum_{i=1}^{k^*} \mathbb{P}(Z > \frac{c_{\mathcal{I}}^2\beta_{\min}^2}{\sigma^2}) = k^*\mathbb{P}(Z > \frac{c_{\mathcal{I}}^2\beta_{\min}^2}{\sigma^2}), \end{aligned} \quad (23)$$

where  $Z = \frac{\|\mathbf{w}\|_2^2}{\sigma^2} \sim \chi_n^2$ . Since,  $Z$  is a B.I.P R.V with distribution independent of  $\sigma^2$  and  $\frac{c_{\mathcal{I}}^2\beta_{\min}^2}{\sigma^2} \rightarrow \infty$  as  $\sigma^2 \rightarrow 0$ , we have  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(Z > \frac{c_{\mathcal{I}}^2\beta_{\min}^2}{\sigma^2}) = 0$ . This implies that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) = 1$ . To summarize, if  $\text{erc}(\mathbf{X}, \mathcal{I}) < 1$  and OMP runs exactly  $k^*$  iterations, then the true support can be detected exactly at high SNR.

The conditional probability  $\mathbb{P}(\mathcal{E}_2|\mathcal{E}_1)$  is given by  $\mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) = \mathbb{P}(\{\text{SC is not satisfied for } i = 1, \dots, k^* - 1\} \cap \{\text{SC is satisfied for } i = k^*\}|\mathcal{E}_1)$ . Complementing and applying union bound gives

$$\begin{aligned} \mathbb{P}(\mathcal{E}_2^C|\mathcal{E}_1) &\leq \sum_{i=1}^{k^*-1} \mathbb{P}(\{\text{SC is satisfied for } i\}|\mathcal{E}_1) + \\ &\quad \mathbb{P}(\{\text{SC is not satisfied for } k^*\}|\mathcal{E}_1). \end{aligned} \quad (24)$$

**Proof of Theorem 10:-** For RPSC, the SC is given by  $\{\|\mathbf{r}^i\|_2 < \sigma\Gamma_4\}$ . First consider  $\mathbb{P}(S_i) = \mathbb{P}(\|\mathbf{r}^i\|_2 < \sigma\Gamma_4)$  for  $i < k^*$  in (24). Using triangle inequality,  $\|\mathbf{r}^i\|_2 \geq \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^i})\mathbf{X}\beta\|_2 - \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^i})\mathbf{w}\|_2$ . Conditioned on  $\mathcal{E}_1$ , we have  $\mathcal{J}^i \subset \mathcal{I}$  for  $i < k^*$  and hence  $\exists \lambda_i > 0$  such that  $\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^i})\mathbf{X}\beta\|_2 > \lambda_i$ , for all  $\sigma^2 > 0$ . Further,  $\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^i})\mathbf{w}\|_2 \leq \|\mathbf{w}\|_2$ . Applying these bounds in  $\mathbb{P}(S_i) = \mathbb{P}(\|\mathbf{r}^i\|_2 < \sigma\Gamma_4)$  gives

$$\mathbb{P}(S_i) \leq \mathbb{P}(\|\mathbf{w}\|_2 + \sigma\Gamma_4 > \lambda_i), \quad \forall i < k^*. \quad (25)$$

Since,  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ , we have  $\|\mathbf{w}\|_2 \xrightarrow{P} 0$  as  $\sigma^2 \rightarrow 0$ . By the hypothesis of Theorem 10,  $\lim_{\sigma^2 \rightarrow 0} \sigma\Gamma_4 = 0$ . Hence,

$\|\mathbf{w}\|_2 + \sigma\Gamma_4 \xrightarrow{P} 0$  as  $\sigma^2 \rightarrow 0$ . Now by the definition of C.I.P,  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\|\mathbf{w}\|_2 + \sigma\Gamma_4 > \lambda_i) = 0$ . This implies that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(S_i) = 0, \forall i < k^*$ .

Next consider  $\mathbb{P}(S_{k^*})$  in (24). Conditioned on  $\mathcal{E}_1$ , all the first  $k^*$  iterations of OMP are correct, i.e.,  $\mathcal{J}^{k^*} = \mathcal{I}$ . This implies that  $\|\mathbf{r}^{k^*}\|_2^2 = \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{y}\|_2^2 = \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_2^2 \sim \sigma^2\chi_{n-k^*}^2$ . Consequently,  $\mathbb{P}(S_{k^*}) = \mathbb{P}(\|\mathbf{r}^{k^*}\|_2^2 > \sigma^2\Gamma_4^2) = \mathbb{P}(Z > \Gamma_4^2)$ ,

where  $Z = \frac{\|\mathbf{r}^{k^*}\|_2^2}{\sigma^2} \sim \chi_{n-k^*}^2$ . Since  $Z$  is a B.I.P R.V with distribution independent of  $\sigma^2$  and  $\Gamma_4 \rightarrow \infty$  as  $\sigma^2 \rightarrow 0$ , it follows that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(S_{k^*}) = 0$ . Substituting

$\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(S_i) = 0$  for  $i \leq k^*$  in (24), we have  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) = 1$ . Combining this with  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) = 1$  gives  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\hat{\mathcal{I}} = \mathcal{I}) = \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) = 1$ . ■

**Proof of Theorem 11:-** For RCSC, the SC is given by  $\{\|\mathbf{X}^T\mathbf{r}^i\|_\infty < \sigma\Gamma_5\}$ . First consider  $\mathbb{P}(S_i) = \mathbb{P}(\|\mathbf{X}^T\mathbf{r}^i\|_\infty < \sigma\Gamma_5)$  for  $i < k^*$  in (24).  $\|\mathbf{X}^T\mathbf{r}^i\|_\infty$  can be lower bounded using triangle inequality as  $\|\mathbf{X}^T\mathbf{r}^i\|_\infty \geq \|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}})\mathbf{X}\beta\|_\infty - \|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}})\mathbf{w}\|_\infty$ . Further,  $\|\mathbf{X}_i\|_2 = 1$  implies that  $\|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}})\mathbf{w}\|_\infty \leq \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^{i-1}})\mathbf{w}\|_2 \leq \|\mathbf{w}\|_2$ . Conditioned on  $\mathcal{E}_1$ , we have  $\mathcal{J}^i \subset \mathcal{I}$  for  $i < k^*$  and hence  $\exists \lambda_i > 0$  such that  $\|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}^i})\mathbf{X}\beta\|_\infty > \lambda_i$ , for all  $\sigma^2 > 0$ . Applying these bounds in  $\mathbb{P}(S_i) = \mathbb{P}(\|\mathbf{X}^T\mathbf{r}^i\|_\infty < \sigma\Gamma_5)$  gives

$$\mathbb{P}(S_i) \leq \mathbb{P}(\|\mathbf{w}\|_2 + \sigma\Gamma_5 > \lambda_i), \quad \forall i < k^*. \quad (26)$$

Following the same arguments used in the proof of Theorem 10 we have  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(S_i) = 0, \forall i < k^*$ .

Next we consider  $\mathbb{P}(S_{k^*}) = \mathbb{P}(\|\mathbf{X}^T\mathbf{r}^{k^*}\|_\infty > \sigma\Gamma_5)$ . Since, the first  $k^*$  iterations are correct, i.e.,  $\mathcal{J}^{k^*} = \mathcal{I}$ , we have  $\|\mathbf{X}^T\mathbf{r}^{k^*}\|_\infty = \|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{y}\|_\infty = \|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_\infty$ . Using Cauchy Schwartz inequality and  $\|\mathbf{X}_j\|_2 = 1$ , it follows that  $\|\mathbf{X}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_\infty \leq \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_2$ . Hence,  $\mathbb{P}(S_{k^*}) \leq \mathbb{P}(\frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_2^2}{\sigma^2} > \Gamma_5^2)$ . Since,  $\frac{\|(\mathbf{I}_n - \mathbf{P}_{\mathcal{I}})\mathbf{w}\|_2^2}{\sigma^2} \sim \chi_{n-k^*}^2$  is a B.I.P R.V and the term  $\Gamma_5^2 \rightarrow \infty$ , it follows that  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(S_{k^*}) = 0$ . Substituting  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(S_i) = 0$  for  $i \leq k^*$  in (24), we have  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) = 1$ . Combining this with

$\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) = 1$  gives  $\lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\hat{\mathcal{I}} = \mathcal{I}) = \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_1) \lim_{\sigma^2 \rightarrow 0} \mathbb{P}(\mathcal{E}_2|\mathcal{E}_1) = 1$ . ■

*Remark 6.* The following observations can be made about the convergence rates in RPSC and RCSC. The rate at which  $\mathbb{P}(\mathcal{E}_1)$  converges to one is independent of  $\Gamma_4$  or  $\Gamma_5$ . First consider  $\mathbb{P}(S_i)$  for  $i < k^*$  and let  $\Gamma_4 = \frac{1}{\sigma^\alpha}$  be the SNR adaptation. Then the rate at which  $\mathbb{P}(S_i)$  converges to zero is maximum when  $\alpha = 0$  and decreases with increasing  $\alpha$ . However, the rate at which  $\mathbb{P}(S_{k^*})$  converges to zero increases with increasing  $\alpha$ .

## VI. NUMERICAL SIMULATIONS

Here we numerically verify the results proved in Theorems 1-11. We consider two classes of matrices for simulations.

**ERC matrix:** We consider a  $n \times 2n$  matrix  $\mathbf{X}$  formed by the concatenation of a  $n \times n$  identity matrix and a Hadamard matrix of size  $n \times n$  denoted by  $\mathbf{H}_n$ , i.e.,  $\mathbf{X} = [\mathbf{I}_n, \mathbf{H}_n]$ . It is well known that this matrix has mutual coherence  $\mu_{\mathbf{X}} = \frac{1}{\sqrt{n}}$  [Chapter 2, [39]]. We fix  $n$  as  $n = 32$  and for this value of

$n$ ,  $\mathbf{X}$  satisfy MIC for any  $\beta$  with sparsity  $k^* \leq \frac{1}{2}(1 + \sqrt{n}) = 3.3284$ . As explained in section II, MIC implies ERC also.

**Random matrix:** A random matrix  $\mathbf{X}$  is generated using *i.i.d*  $\mathbf{X}_{i,j} \sim \mathcal{N}(0, 1)$  R.Vs and columns in this matrix are later normalised to have unit  $l_2$  norm. In each iteration the matrix  $\mathbf{X}$  is independently generated. The matrix support pair thus generated in each iteration may or may not satisfy ERC.

All non zero entries have same magnitude (denoted by  $\beta_k$  in figures) but random signs. Further, the  $k^*$  non zero entries are selected randomly from the set  $[p]$ . The figures are produced after performing  $10^5$  iterations at each SNR level.

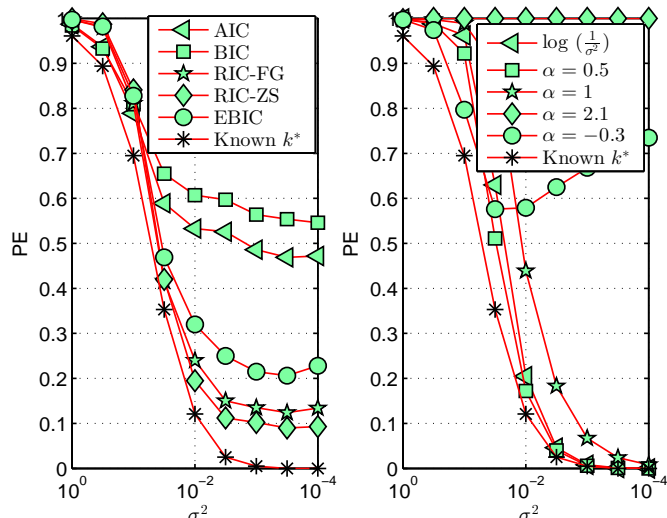


Fig. 1. Performance of  $l_0$ -penalty with a  $5 \times 10$  random matrix.  $\beta_k = \pm 1$  and  $k^* = 2$ .

### A. Performance of $l_0$ -penalty.

The performance of  $l_0$ -penalty with different values of  $\Gamma_0$  is reported in Fig.1. The matrix under consideration is a  $5 \times 10$  random matrix. “Known  $k^*$ ” represents the performance of an oracle SSP with *a priori* information of  $k^*$ . This SSP estimates  $\hat{\mathcal{I}}$  using  $\hat{\mathcal{I}} = \arg \min_{\mathcal{J} \subset [p], |\mathcal{J}|=k^*} \|(\mathbf{I}_n - \mathbf{P}_{\mathcal{J}})\mathbf{y}\|_2^2$  and will have superior performance when compared with  $l_0$ -penalty which is oblivious to  $k^*$ .

L.H.S of Fig.1 gives the performance of  $l_0$ -penalty with SNR independent values of  $\Gamma_0$  discussed in literature. AIC uses  $\Gamma_0 = 2$ , BIC uses  $\Gamma_0 = \log(n)$ , RIC-FG uses  $\Gamma_0 = 2 \log(p)$  [29], RIC-ZS uses  $\Gamma_0 = 2 \log(p) + 2 \log(\log(p))$  [30] and EBIC uses  $\Gamma_0 = \log(n) + \frac{2}{\|\mathbf{b}\|_0} \log(\binom{p}{\|\mathbf{b}\|_0})$ . As predicted by Theorem 2,  $l_0$ -penalty with all these values of  $\Gamma_0$  are inconsistent at high SNR. The performance of RIC-ZS is the best among the values of  $\Gamma_0$  under consideration. The performance of BIC and AIC are much poorer compared to other schemes. When  $n = 5$ ,  $\Gamma_0 = 2$  in AIC is bigger than  $\Gamma_0 = \log(n)$  of BIC and this explains the inferior performance of BIC *viz a viz* AIC. For higher values of  $n$ , BIC will perform better than AIC.

R.H.S gives the performance of  $l_0$ -penalty with  $\Gamma_0 = f(\sigma^2)[\log(n) + \frac{2}{\|\mathbf{b}\|_0} \log(\binom{p}{\|\mathbf{b}\|_0})]$ , i.e., a SNR adaptation is added to EBIC penalty. “ $\log(\frac{1}{\sigma^2})$ ” in Fig.1 represents  $f(\sigma^2) = \log(\frac{1}{\sigma^2})$ . This SNR adaptation satisfies the conditions in Theorem 1 and is common in popular MOS criteria like NMDL, g-MDL etc. [17]. The schemes represented using  $\alpha = (\cdot)$  has  $f(\sigma^2) = \frac{1}{\sigma^\alpha}$ . Among the values of  $\alpha$  considered in Fig.1,  $\alpha = 0.5$  and  $\alpha = 1$  satisfies the conditions in Theorem 1,  $\alpha = -0.3$  violates Theorem 2 and  $\alpha = 2.1$  violates Theorem 3 respectively. As predicted by Theorems 1-3, only “ $\log(\frac{1}{\sigma^2})$ ”,  $\alpha = 0.5$  and  $\alpha = 1$  that satisfies the conditions in Theorem 1 are high SNR consistent. This verify

the NSCs derived in section III. Further, the performance of  $l_0$ -penalty with  $\Gamma_0$  represented by “ $\log(\frac{1}{\sigma^2})$ ” and  $\alpha = 0.5$  are very close to the optimal scheme represented by “Known  $k^*$ ” across the entire SNR range. This suggest the finite SNR utility of the SNR adaptations suggested by Theorem 1.

### B. Performance of $l_1$ -penalty and $l_1$ -error at high SNR.

L.H.S of Fig.2 gives the performance of  $l_1$ -penalty and R.H.S of Fig.2 gives the performance of  $l_1$ -error respectively. Both these SSPs are evaluated for the  $32 \times 64$  ERC matrix previously defined and a  $75 \times 100$  random matrix. “ $2\sqrt{2 \log(p)}$ ” in L.H.S represents the performance of  $l_1$ -penalty with  $\Gamma_1 = 2\sqrt{2 \log(p)}$  [34] and “ $\alpha = (\cdot)$ ” represents  $l_1$ -penalty with  $\Gamma_1 = \frac{1}{\sigma^\alpha} 2\sqrt{2 \log(p)}$ . Similarly, in the R.H.S, “ $\sqrt{n + 2\sqrt{2n}}$ ” represents the  $l_1$ -error with  $\Gamma_2 = \sqrt{n + 2\sqrt{2n}}$  [36] and “ $\alpha = (\cdot)$ ” represents  $l_1$ -error with  $\Gamma_2 = \frac{1}{\sigma^\alpha} \sqrt{n + 2\sqrt{2n}}$ . In both cases,  $\alpha = (\cdot)$  incorporates a SNR adaptation into a well known value of  $\Gamma_1$  and  $\Gamma_2$ . By Theorems 4-8, these SNR adaptations are consistent iff  $0 < \alpha < 1$ .

First we consider the performance of  $l_1$ -penalty for the matrix  $\mathbf{X}$  satisfying ERC. It is clear from Fig.2 that  $l_1$ -penalty with  $\Gamma_1 = 2\sqrt{2 \log(p)}$  floors at high SNR with a  $PE \approx 10^{-2.5}$ . Hence,  $l_1$ -penalty with  $\Gamma_1 = 2\sqrt{2 \log(p)}$  is inconsistent at high SNR and this validates Theorem 5. Other  $\sigma^2$  independent values of  $\Gamma_1$  discussed in Section IV also floors at high SNR. On the contrary,  $l_1$ -penalty with SNR dependent  $\Gamma_1$  does not floor at high SNR and this validates Theorem 4. Further,  $\Gamma_1$  with  $\alpha = 0.1$  performs better than  $\Gamma_1 = 2\sqrt{2 \log(p)}$  even for  $\sigma^2 \approx 0.01$ . In the same setting,  $l_1$ -error with  $\Gamma_2 = \sqrt{n + 2\sqrt{2n}}$  is inconsistent at high SNR. In fact PE for  $\Gamma_2 = \sqrt{n + 2\sqrt{2n}}$  floors at  $PE \approx 10^{-1.25}$  at high SNR. It is evident from Fig.2 that  $\Gamma_2 = \frac{1}{\sigma^\alpha} \sqrt{n + 2\sqrt{2n}}$ , where  $\alpha = 0.15$  and  $\alpha = 0.3$  are high SNR consistent. These results validates Theorems 7-8. In fact  $l_1$ -error with  $\alpha = 0.15$  and  $\alpha = 0.3$  performs much better than the SNR independent  $\Gamma_2 = \sqrt{n + 2\sqrt{2n}}$  from  $\sigma^2 \approx 0.01$  onwards.

Next we consider the performance of  $l_1$ -penalty and  $l_1$ -error when  $\mathbf{X}$  is a random  $75 \times 100$  matrix. Here also  $l_1$ -penalty and  $l_1$ -error with values of  $\Gamma_1$  and  $\Gamma_2$  independent of  $\sigma^2$  floors at high SNR. However, unlike the case of ERC matrix,  $\Gamma_1$  and  $\Gamma_2$  with SNR adaptations stipulated by Theorem 4 and Theorem 7 appears to floor at high SNR. This is because of the fact that there is a non zero probability  $p_{erc} > 0$  with which a particular realization of  $(\mathbf{X}, \beta)$  pair fails to satisfy conditions like ERC. In fact  $p_{erc}$  decreases exponentially with increasing  $n$ . Hence, for random matrices  $p_{erc}$  dictates the PE at which  $l_1$ -penalty and  $l_1$ -error floors. Note that the level at which PE of  $l_1$ -penalty with  $\Gamma_1$  and  $\Gamma_2$  satisfying the SNR adaptations stipulated by Theorem 4 and Theorem 7 floors is significantly lower than the case with SNR independent  $\Gamma_1$  and  $\Gamma_2$ . This indicates that the proposed SNR adaptations can improve performance in situations beyond the regression classes for which high SNR consistency is established.



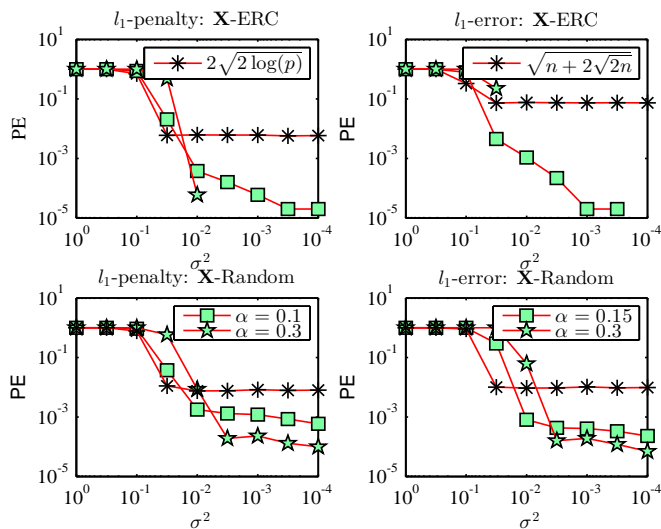


Fig. 2. Performance of  $l_1$ -penalty and  $l_1$ -error for a  $32 \times 64$  ERC matrix and  $75 \times 100$  random matrix.  $k^* = 3$  and  $\beta_k = \pm 1$ .

### C. Performance of OMP at high SNR.

L.H.S of Fig.3 presents the performance of OMP with RPSC and R.H.S presents the performance of OMP with RCSC respectively. Both these SSPs are evaluated for the ERC matrix previously defined and a  $75 \times 100$  random matrix. “Known  $k^*$ ” represents a hypothetical SSP which runs OMP for exactly  $k^* = 3$  iterations. “ $f_n$ ” in the L.H.S represents the performance of RPSC with  $\Gamma_4 = \sqrt{n + 2\sqrt{n \log(n)}}$  and “ $\alpha = (\cdot)$ ” represents the performance of RPSC with  $\Gamma_4 = \frac{1}{\sigma^\alpha} \sqrt{n + 2\sqrt{n \log(n)}}$ . Similarly, “ $f_p$ ” in the R.H.S represents the performance of RCSC with  $\Gamma_5 = \sqrt{4 \log(p)}$  and “ $\alpha = (\cdot)$ ” represents the performance of RCSC with  $\Gamma_5 = \frac{1}{\sigma^\alpha} \sqrt{4 \log(p)}$ .

$\Gamma_4 = \sqrt{n + 2\sqrt{n \log(n)}}$  and  $\Gamma_5 = \sqrt{4 \log(p)}$  are suggested in [23]. “ $\alpha = (\cdot)$ ” in both cases incorporate a SNR adaptation into these well known stopping parameters. It is clear from the Fig.3 that OMP with SC independent of  $\sigma^2$  floors at high SNR for both ERC and random matrices, whereas, the flooring of PE is not present in OMP with SC satisfying Theorems 10 and 11 for ERC matrix. For random matrix, the performance of OMP with proposed SNR adaptations floors at a PE level equal to that of OMP with known  $k^*$ . This flooring is also due to the causes explained for  $l_1$ -penalty and  $l_1$ -error.

### D. On the choice of SNR adaptations.

Fig.4 presents the effect of signal strength  $|\beta_j|$  and SNR adaptations on the convergence rates of  $l_1$ -penalty and OMP-RPSC. “ $f_n$ ” represents RPSC with  $\Gamma_4 = \sqrt{n + 2\sqrt{2 \log(n)}}$  as before. “ $\alpha = (\cdot)$ ” represents  $l_1$ -penalty with  $\Gamma_1 = \frac{1}{\sigma^\alpha} 2\sqrt{2 \log(p)}$  and RPSC with  $\Gamma_4 = \frac{1}{\sigma^\alpha} \sqrt{n + 2\sqrt{2 \log(n)}}$ . By Theorems 4 and 10, the SNR adaptations represented by  $\alpha = (\cdot)$  will be consistent for both  $l_1$ -penalty and RPSC iff  $0 < \alpha < 1$ . However, the deviations from the base tuning parameters (i.e.,  $2\sqrt{2 \log(p)}$  and  $\sqrt{n + 2\sqrt{2 \log(n)}}$ ) will be

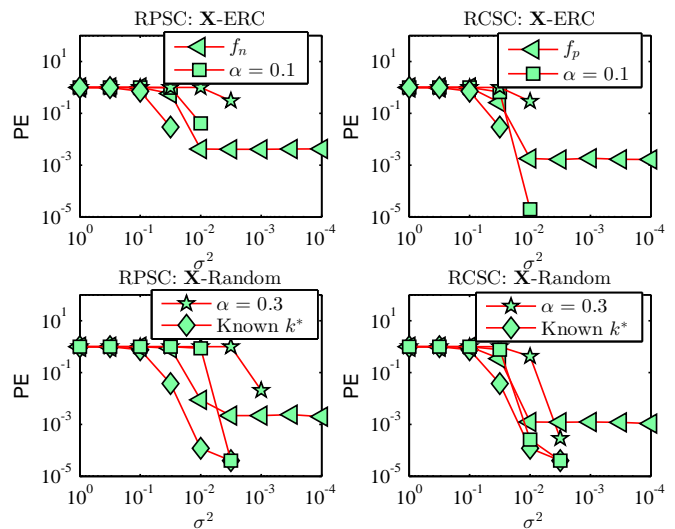


Fig. 3. Performance of OMP with RPSC and RCSC for a  $32 \times 64$  ERC matrix and  $75 \times 100$  random matrix.  $k^* = 3$  and  $\beta_k = \pm 1$ .

more pronounced as  $\alpha$  increases. This will influence the rate at which  $PE$  converges to zero.

At very high SNR, the performance of  $l_1$ -penalty and OMP-RPSC with larger values of  $\alpha$  will be better. This is true for both low and high values of regression coefficients (i.e.,  $\beta_j = 0.5$  and  $\beta_j = 3$ ). Throughout the moderate to high SNR range, the performance of these algorithms with high values of  $\alpha$  will be poor in comparison with the base tuning parameter when  $|\beta_j|$  is low. In the same SNR and signal strength regime the performance with low values of  $\alpha$  will be better than both base tuning parameter and high value of  $\alpha$ . As the signal strength improves, the performance of these algorithms improves for all values of  $\alpha$ . However, the performance with high values of  $\alpha$  will be much better than the performance with low values of  $\alpha$  when  $|\beta_j|$  is high. Note that the PE with base tuning parameter floors at the same value irrespective of signal strength. The numerical results are in line with the inferences derived from the convergence rate analysis of  $l_1$ -penalty. Similar inferences can be derived from the numerical experiments (not shown) conducted for other CS algorithms considered in this paper.

Note that the very high SNR regime is rarely encountered in practice. Further, a low value of  $\alpha$  will provide a performance atleast as good as the performance of the base parameter in the moderate SNR range irrespective of the signal strength and a progressively improving performance as the SNR or the signal strength improves. Hence, by following the philosophy of minimizing the worst case risk, it will be advisable to choose smaller values of  $\alpha$  like  $\alpha = 0.1$  for practical applications.

## VII. CONCLUSION

NSCs for the high SNR consistency of CS algorithms like  $l_0$ -penalty,  $l_1$ -penalty,  $l_1$ -error, DS and OMP are derived in this paper. Aforementioned algorithms with the tuning parameters discussed in literature are analytically and numerically shown to be inconsistent at high SNR. Novel tuning parameters for these CS algorithms are derived based on the sufficient conditions and justified using convergence rate analysis. CS

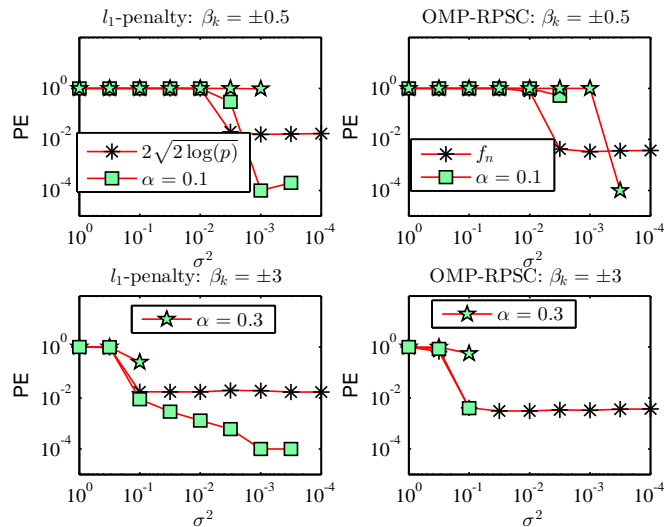


Fig. 4. Convergence rates for OMP with RPSC and  $l_1$ -penalty for a  $32 \times 64$  ERC matrix and  $k^* = 3$ .

algorithms with the proposed tuning parameters are numerically shown to perform better than existing tuning parameters.

## REFERENCES

- [1] Y. C. Eldar and G. Kutyniok, *Compressed sensing: Theory and applications*. Cambridge University Press, 2012.
- [2] T. T. Emmanuel Candes, "The Dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ ," *Ann. Stat.*, vol. 35, no. 6, pp. 2313–2351, 2007.
- [3] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.
- [4] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, March 2006.
- [5] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.
- [6] M. Masood and T. Y. Al-Naffouri, "Sparse reconstruction using distribution agnostic Bayesian matching pursuit," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5298–5309, 2013.
- [7] Z. Ben-Haim, Y. C. Eldar, and M. Elad, "Coherence-based performance guarantees for estimating a sparse vector under random noise," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5030–5043, 2010.
- [8] K. Sreejith and S. Kalyani, "High SNR consistent thresholding for variable selection," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 1940–1944, Nov 2015.
- [9] P. Zhao and B. Yu, "On model selection consistency of LASSO," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, 2006.
- [10] Q. Ding and S. Kay, "Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1959–1969, May 2011.
- [11] S. Kay, "Exponentially embedded families - New approaches to model order estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, no. 1, pp. 333–345, Jan 2005.
- [12] J. Rissanen, "MDL denoising," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2537–2543, Nov 2000.
- [13] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *J. Amer. Statist. Assoc.*, vol. 96, no. 454, pp. 746–774, 2001.
- [14] P. Stoica and P. Babu, "On the proper forms of BIC for model order selection," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4956–4961, Sept 2012.
- [15] J. Nielsen, M. Christensen, and S. Jensen, "Bayesian model comparison and the BIC for regression models," in *ICASSP*, May 2013, pp. 6362–6366.
- [16] J. Rissanen, T. Roos, and P. Myllymki, "Model selection by sequentially normalized least squares," *J. Multivariate Anal.*, vol. 101, no. 4, pp. 839–849, 2010.
- [17] S. Kallummil and S. Kalyani, "High SNR consistent linear model order selection and subset selection," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4307–4322, Aug 2016.
- [18] D. Schmidt and E. Makalic, "The consistency of MDL for linear regression models with increasing signal-to-noise ratio," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1508–1510, March 2012.
- [19] J. Määttä, D. F. Schmidt, and T. Roos, "Subset selection in linear regression using sequentially normalized least squares: Asymptotic theory," *SCAND. J. STAT.*, 2015.
- [20] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Springer, 2013.
- [21] J. Wang, "Support recovery with orthogonal matching pursuit in the presence of noise," *IEEE Trans. Signal Process.*, vol. 63, no. 21, pp. 5868–5877, Nov 2015.
- [22] J. Wang and B. Shim, "On the recovery limit of sparse signals using orthogonal matching pursuit," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4973–4976, 2012.
- [23] T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4680–4688, July 2011.
- [24] S. Kay, Q. Ding, B. Tang, and H. He, "Probability density function estimation using the EEF with application to subset/feature selection," *IEEE Trans. Signal Process.*, vol. 64, no. 3, pp. 641–651, Feb 2016.
- [25] B. Shim and B. Song, "Multiuser detection via compressive sensing," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 972–974, July 2012.
- [26] A. K. Fletcher, S. Rangan, and V. K. Goyal, "On-off random access channels: A compressed sensing framework," *arXiv preprint arXiv:0903.1022*, 2009.
- [27] P. Gerstoft, A. Xenaki, and C. F. Mecklenbräuer, "Multiple and single snapshot compressive beamforming," *The Journal of the Acoustical Society of America*, vol. 138, no. 4, pp. 2003–2014, 2015.
- [28] K. L. Chung, *A course in probability theory*. Academic press, 2001.
- [29] D. P. Foster and E. I. George, "The risk inflation criterion for multiple regression," *Ann. Stat.*, pp. 1947–1975, 1994.
- [30] Y. Zhang and X. Shen, "Model selection procedure for high-dimensional data," *Stat. Anal. Data Min.*, vol. 3, no. 5, pp. 350–358, 2010.
- [31] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.
- [32] Y. Kim, S. Kwon, and H. Choi, "Consistent model selection criteria on high dimensions," *J. Mach. Learn. Res.*, vol. 13, no. Apr, pp. 1037–1057, 2012.
- [33] X. Yan and X. Su, *Linear regression analysis: Theory and computing*. World Scientific, 2009.
- [34] E. J. Candès, Y. Plan *et al.*, "Near-ideal model selection by  $l_1$  minimization," *Ann. Stat.*, vol. 37, no. 5A, pp. 2145–2177, 2009.
- [35] E. J. Candès and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7235–7254, 2011.
- [36] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.
- [37] T. T. Cai, G. Xu, and J. Zhang, "On recovery of sparse signals via  $l_1$  minimization," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3388–3397, July 2009.
- [38] T. T. Cai, L. Wang, and G. Xu, "Stable recovery of sparse signals and an oracle inequality," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3516–3522, 2010.
- [39] M. Elad, *Sparse and redundant representation*. Springer, 2010.