Signal and Noise Statistics Oblivious Orthogonal Matching Pursuit

Sreejith Kallummil [*1] **Sheetal Kalyani** [*2]

Abstract

Orthogonal matching pursuit (OMP) is a widely used algorithm for recovering sparse high dimensional vectors in linear regression models. The optimal performance of OMP requires *a priori* knowledge of either the sparsity of regression vector or noise statistics. Both these statistics are rarely known *a priori* and are very difficult to estimate. In this paper, we present a novel technique called residual ratio thresholding (RRT) to operate OMP without any *a priori* knowledge of sparsity and noise statistics and establish finite sample and large sample support recovery guarantees for the same. Both analytical results and numerical simulations in real and synthetic data sets indicate that RRT has a performance comparable to OMP with *a priori* knowledge of sparsity and noise statistics.

1. Introduction

This article deals with the estimation of the regression vector $\boldsymbol{\beta} \in \mathbb{R}^p$ in the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a known design matrix with unit Euclidean norm columns, $\mathbf{w}$ is the noise vector and $\mathbf{y}$ is the observation vector. Throughout this article, we assume that the entries of the noise $\mathbf{w}$ are independent, zero mean and Gaussian distributed with variance σ^2. We consider the high dimensional and sample starved scenario of $n < p$ or $n \ll p$ where classical techniques like ordinary least squares (OLS) are no longer applicable. This problem of estimating high dimensional vectors in sample starved scenarios is ill-posed even in the absence of noise unless strong structural assumptions are made on $\mathbf{X}$ and $\boldsymbol{\beta}$. A widely used and practically valid assumption is sparsity. The vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is sparse if the support of $\boldsymbol{\beta}$ given by $\mathcal{S} = supp(\boldsymbol{\beta}) = \{k : \boldsymbol{\beta}_k \neq 0\}$ has cardinality $k_0 = card(\mathcal{S}) \ll p$.

A number of algorithms like least absolute shrinkage and selection operator (LASSO)(Tropp, 2006; Tibshirani, 1996), Dantzig selector (DS)(Candes & Tao, 2007), subspace pursuit (SP)(Dai & Milenkovic, 2009), OMP (Pati et al., 1993; Mallat & Zhang, 1993; Tropp, 2004; Cai & Wang, 2011), elastic net (Zou & Hastie, 2005) etc. are proposed to efficiently estimate $\boldsymbol{\beta}$. Tuning the hyper parameters of aforementioned algorithms to achieve optimal performance require *a priori* knowledge of signal parameters like sparsity k_0 or noise statistics like σ^2 etc. Unfortunately, these parameters are rarely known *a priori*. To the best of our knowledge, no computationally efficient technique to estimate k_0 is reported in open literature. However, limited success on the estimation of σ^2 has been reported in literature (Dicker, 2014; Fan et al., 2012; Dicker & Erdogdu, 2016; Bayati et al., 2013). However, the performance of these σ^2 estimates when used for tuning hyper parameters in LASSO, DS, OMP etc. are largely unknown. Generalised techniques for hyper parameter selection like cross validation (CV)(Arlot et al., 2010), re-sampling (Meinshausen & Bühlmann, 2010) etc. are computationally challenging. Further, CV is reported to have poor variable selection behaviour(Chichignoud et al., 2016; Arlot et al., 2010). Indeed, algorithms that are oblivious to signal and noise statistics are also proposed in literature. This include algorithms inspired or related to LASSO like square root LASSO(Belloni et al., 2011), AV_∞ (Chichignoud et al., 2016), approximate message passing (Mousavi et al., 2013; Bayati et al., 2013) etc. and ridge regression inspired techniques like least squares adaptive thresholding (LAT), ridge adaptive thresholding (RAT)(Wang et al., 2016) etc. However, most of existing signal and noise statistics oblivious sparse recovery techniques have only large sample performance guarantees. Further, many of these techniques assume that design matrix $\mathbf{X}$ is sampled from a random ensemble, a condition which is rarely satisfied in practice.

1.1. Contributions of this paper

This article present a novel technique called residual ratio thresholding (RRT) for finding a "good" estimate of support $\mathcal{S}$ from the data dependent/adaptive sequence of supports generated by OMP. RRT is analytically shown to accomplish exact support recovery, (i.e., identifying $\mathcal{S}$) under the same finite sample and deterministic constraints on $\mathbf{X}$ like

[*]Equal contribution [1]Department of Electrical Engineering, IIT Madras, India [2]Department of Electrical Engineering, IIT Madras, India. Correspondence to: Sreejith Kallummil <sreejith.k.venugopal@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

restricted isometry constants (RIC) or mutual coherence required by OMP with *a priori* knowledge of k_0 or σ^2. However, the signal to noise ratio (SNR=$\|\mathbf{X}\boldsymbol{\beta}\|_2^2/n\sigma^2$) required for support recovery using RRT is slightly higher than that of OMP with *a priori* knowledge of k_0 or σ^2. This extra SNR requirement is shown to decrease with the increase in sample size n. RRT and OMP with *a priori* knowledge of k_0 or σ^2 are shown to be equivalent as $n \to \infty$ in terms of the SNR required for support recovery. RRT involves a tuning parameter α that can be set independent of ambient SNR or noise statistics. The hyper parameter α in RRT have an interesting semantic interpretation of being the high SNR upper bound on support recovery error. Also RRT is asymptotically tuning free in the sense that a very wide range of α deliver similar performances as $n \to \infty$. Numerical simulations indicate that RRT can deliver a highly competitive performance when compared to OMP having *a priori* knowledge of k_0 or σ^2, OMP with k_0 estimated using CV and the recently proposed LAT algorithm. Further, RRT also delivered a highly competitive performance when applied to identify outliers in real data sets, an increasingly popular application of sparse estimation algorithms(Mitra et al., 2010; 2013).

The remainder of this article is organised as follows. In section 2 we discuss OMP algorithm. RRT algorithm is presented in Section 3. Section 4 presents theoretical performance guarantees for RRT. Section 5 presents numerical simulation results. All the proofs are provided in the supplementary material.

1.2. Notations used

$\|\mathbf{x}\|_q = \left(\sum_{k=1}^{p} |\mathbf{x}_k|^q\right)^{\frac{1}{q}}$ is the l_q norm of $\mathbf{x} \in \mathbb{R}^p$. $\mathbf{0}_n$ is the $n \times 1$ zero vector and $\mathbf{I}_n$ is the $n \times n$ identity matrix. $span(\mathbf{X})$ is the column space of $\mathbf{X}$. $\mathbf{X}^\dagger = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the Moore-Penrose pseudo inverse of $\mathbf{X}$. $\mathbf{X}_\mathcal{J}$ denotes the sub-matrix of $\mathbf{X}$ formed using the columns indexed by $\mathcal{J}$. $\mathcal{N}(\mathbf{u}, \mathbf{C})$ represents a Gaussian random vector (R.V) with mean $\mathbf{u}$ and covariance matrix $\mathbf{C}$. $\mathbb{B}(a, b)$ denotes a Beta R.V with parameters a and b. $\mathbf{a} \sim \mathbf{b}$ implies that $\mathbf{a}$ and $\mathbf{b}$ are identically distributed. $[p]$ represents the floor operator. ϕ represents the null set. For any two sets $\mathcal{J}_1$ and $\mathcal{J}_2$, $\mathcal{J}_1/\mathcal{J}_2$ denotes the set difference. $\mathbf{a} \xrightarrow{P} \mathbf{b}$ represents the convergence of R.V $\mathbf{a}$ to R.V $\mathbf{b}$ in probability.

2. Orthogonal Matching Pursuit (OMP)

OMP (Algorithm 1) starts with a null support estimate and in each iteration it adds that column index to the current support which is the most correlated with the previous residual $\mathbf{r}^{k-1}$, i.e., $t_k = \arg\max_j |\mathbf{X}_j^T\mathbf{r}^{k-1}|$. Then a LS estimate of $\boldsymbol{\beta}$ restricted to the current support $\mathcal{S}_{omp}^k$ is computed as an intermediate estimate of $\boldsymbol{\beta}$ and this estimate is used to update the residual. Note that $\mathbf{P}_k$ in Algorithm 1 refers to $\mathbf{X}_{\mathcal{S}_{omp}^k}\mathbf{X}_{\mathcal{S}_{omp}^k}^\dagger$, the projection matrix onto $span(\mathbf{X}_{\mathcal{S}_{omp}^k})$. Since the residual $\mathbf{r}^k$ is orthogonal to $span(\mathbf{X}_{\mathcal{S}_{omp}^k})$, $\mathbf{X}_j^T\mathbf{r}^k = 0$ for all $j \in \mathcal{S}_{omp}^k$. Consequently, $t_{k+1} \notin \mathcal{S}_{omp}^k$, i.e., the same index will not be selected in two different iterations. Hence, $\mathcal{S}_{omp}^{k+1} \supset \mathcal{S}_{omp}^k$, i.e. the support sequence is monotonically increasing. The monotonicity of $\mathcal{S}_{omp}^k$ in turn implies that the residual norm $\|\mathbf{r}^k\|_2$ is a non increasing function of k, i.e, $\|\mathbf{r}^{k+1}\|_2 \leq \|\mathbf{r}^k\|_2$.

Algorithm 1 Orthogonal matching pursuit

Input: Observation $\mathbf{y}$, matrix $\mathbf{X}$
Initialize $\mathcal{S}_0^{omp} = \phi$. $k = 1$ and residual $\mathbf{r}^0 = \mathbf{y}$
repeat
  Identify the next column $t_k = \arg\max_j |\mathbf{X}_j^T\mathbf{r}^{k-1}|$
  Expand current support $\mathcal{S}_{omp}^k = \mathcal{S}_{omp}^{k-1} \cup t_k$
  Restricted LS estimate: $\hat{\boldsymbol{\beta}}_{\mathcal{S}_{omp}^k} = \mathbf{X}_{\mathcal{S}_{omp}^k}^\dagger\mathbf{y}$.
  $\hat{\boldsymbol{\beta}}_{\{1,\ldots,p\}/\mathcal{S}_{omp}^k} = \mathbf{0}_{p-k}$.
  Update residual: $\mathbf{r}^k = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{P}_k)\mathbf{y}$.
  Increment $k \leftarrow k + 1$.
until stopping condition (SC) is true
Output: Support estimate $\hat{S} = \mathcal{S}_{omp}^k$. Vector estimate $\hat{\boldsymbol{\beta}}$

Most of the theoretical properties of OMP are derived assuming *a priori* knowledge of true sparsity level k_0 in which case OMP stops after exactly k_0 iterations(Tropp, 2004; Wang, 2015). When k_0 is not known, one has to rely on stopping conditions (SC) based on the properties of the residual $\mathbf{r}^k$ as k varies. For example, one can stop OMP iterations once the residual power is too low compared to the expected noise power. Mathematically, when the noise $\mathbf{w}$ is l_2 bounded, i.e., $\|\mathbf{w}\|_2 \leq \epsilon_2$ for some *a priori* known ϵ_2, then OMP can be stopped if $\|\mathbf{r}^k\|_2 \leq \epsilon_2$. For a Gaussian noise vector $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$, $\epsilon_\sigma = \sigma\sqrt{n + 2\sqrt{n\log(n)}}$ satisfies(Cai & Wang, 2011)

$$\mathbb{P}(\|\mathbf{w}\|_2 \leq \epsilon_\sigma) \geq 1 - \frac{1}{n}, \tag{1}$$

i.e., Gaussian noise is l_2 bounded with a very high probability. Consequently, one can stop OMP iterations in Gaussian noise once $\|\mathbf{r}^k\|_2 \leq \epsilon_\sigma$.

A number of deterministic recovery guarantees are proposed for OMP. Among these guarantees the conditions based on RIC are the most popular. RIC of order j denoted by δ_j is defined as the smallest value of δ such that

$$(1 - \delta)\|\mathbf{b}\|_2^2 \leq \|\mathbf{X}\mathbf{b}\|_2^2 \leq (1 + \delta)\|\mathbf{b}\|_2^2 \tag{2}$$

hold true for all $\mathbf{b} \in \mathbb{R}^p$ with $\|\mathbf{b}\|_0 = card(supp(\mathbf{b})) \leq j$. A smaller value of δ_j implies that $\mathbf{X}$ act as a near orthogonal

matrix for all j sparse vectors $\mathbf{b}$. Such a situation is ideal for the recovery of a j-sparse vector $\mathbf{b}$ using any sparse recovery technique. The latest RIC based support recovery guarantee using OMP is given in Lemma 1(Liu et al., 2017).

Lemma 1. *OMP with k_0 iterations or SC $\|\mathbf{r}^k\|_2 \leq \|\mathbf{w}\|_2$ can recover any k_0 sparse vector $\boldsymbol{\beta}$ provided that $\delta_{k_0+1} < 1/\sqrt{k_0+1}$ and $\|\mathbf{w}\|_2 \leq \epsilon_{omp} = \boldsymbol{\beta}_{min}\sqrt{1-\delta_{k_0+1}}\left[\frac{1-\sqrt{k_0+1}\delta_{k_0+1}}{1+\sqrt{1-\delta_{k_0+1}^2}-\sqrt{k_0+1}\delta_{k_0+1}}\right]$.*

Since $\mathbb{P}(\|\mathbf{w}\|_2 < \epsilon_\sigma) \geq 1 - 1/n$ when $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$, it follows from Lemma 1 that OMP with k_0 iterations or SC $\|\mathbf{r}^k\|_2 \leq \epsilon_\sigma$ can recover any k_0-sparse vector $\boldsymbol{\beta}$ with probability greater than $1 - 1/n$ provided that $\delta_{k_0+1} < 1/\sqrt{k_0+1}$ and $\epsilon_\sigma \leq \epsilon_{omp}$. Lemma 1 implies that OMP with *a priori* knowledge of k_0 or σ^2 can recover support $\mathcal{S}$ once the matrix satisfies the regularity condition $\delta_{k_0+1} < 1/\sqrt{k_0+1}$ and the SNR is high. It is also known that this RIC condition is worst case necessary. Consequently, Lemma 1 is one of the best deterministic guarantee for OMP available in literature. Note that the mutual incoherence condition given by $\mu_{\mathbf{X}} = \max_{j\neq k}|\mathbf{X}_j^T\mathbf{X}_k| < 1/(2k_0-1)$ also ensures exact support recovery at high SNR. Note that the *a priori* knowledge of k_0 or σ^2 required to materialise the recovery guarantees in Lemma 1 are not available in practical problems. Further, k_0 and σ^2 are very difficult to estimate. This motivates the proposed RRT algorithm which does not require *a priori* knowledge of k_0 or σ^2.

3. Residual Ratio Thresholding (RRT)

RRT is a novel signal and noise statistics oblivious technique to estimate the support $\mathcal{S}$ based on the behaviour of the residual ratio statistic $RR(k) = \|\mathbf{r}^k\|_2/\|\mathbf{r}^{k-1}\|_2$ as k increases from $k = 1$ to a predefined value $k = k_{max} > k_0$. As aforementioned, identifying the support using the behaviour of $\|\mathbf{r}^k\|_2$ requires *a priori* knowledge of σ^2. However, as we will show in this section, support detection using $RR(k)$ does not require *a priori* knowledge of σ^2. Since the residual norms are non negative and non increasing, $RR(k)$ always satisfy $0 \leq RR(k) \leq 1$.

3.1. Minimal superset and implications

Consider running $k_{max} > k_0$ iterations of OMP and let $\{\mathcal{S}_{omp}^k\}_{k=1}^{kmax}$ be the support sequence generated by OMP. Recall that $\mathcal{S}_{omp}^k$ is monotonically increasing.

Definition 1:- The minimal superset in the OMP support sequence $\{\mathcal{S}_{omp}^k\}_{k=1}^{kmax}$ is given by $\mathcal{S}_{omp}^{k_{min}}$, where $k_{min} = \min(\{k : \mathcal{S} \subseteq \mathcal{S}_{omp}^k\})$. When the set $\{k : \mathcal{S} \subseteq \mathcal{S}_{omp}^k\} = \phi$, we set $k_{min} = \infty$ and $\mathcal{S}_{omp}^{k_{min}} = \phi$.

In words, minimal superset is the smallest superset of support $\mathcal{S}$ present in a particular realization of the support estimate sequence $\{\mathcal{S}_{omp}^k\}_{k=1}^{kmax}$. Note that both k_{min} and $\mathcal{S}_{omp}^{k_{min}}$ are unobservable random variables. Since $card(\mathcal{S}_{omp}^k) = k$, $\mathcal{S}_{omp}^k$ for $k < k_0$ cannot satisfy $\mathcal{S} \subseteq \mathcal{S}_{omp}^k$ and hence $k_{min} \geq k_0$. Further, the monotonicity of $\mathcal{S}_{omp}^k$ implies that $\mathcal{S} \subset \mathcal{S}_{omp}^k$ for all $k \geq k_{min}$.

Case 1:- When $k_{min} = k_0$, then $\mathcal{S}_{omp}^{k_0} = \mathcal{S}$ and $\mathcal{S}_{omp}^k \supset \mathcal{S}$ for $k \geq k_0$, i.e., $\mathcal{S}$ is present in the solution path. Further, when $k_{min} = k_0$, it is true that $\mathcal{S}_{omp}^k \subseteq \mathcal{S}$ for $k \leq k_0$.

Case 2:- When $k_0 < k_{min} \leq k_{max}$, then $\mathcal{S}_{omp}^k \neq \mathcal{S}$ for all k and $\mathcal{S}_k^{omp} \supset \mathcal{S}$ for $k \geq k_{min}$, i.e., $\mathcal{S}$ is not present in the solution path. However, a superset of $\mathcal{S}$ is present.

Case 3:- When $k_{min} = \infty$, then $\mathcal{S}_{omp}^k \not\supseteq \mathcal{S}$ for all k, i.e., neither $\mathcal{S}$ nor a superset of $\mathcal{S}$ is present in $\{\mathcal{S}_{omp}^k\}_{k=1}^{kmax}$.

To summarize, exact support recovery using any OMP based scheme including the signal and noise statistics aware schemes is possible only if $k_{min} = k_0$. Whenever $k_{min} > k_0$, it is possible to estimate true support $\mathcal{S}$ without having any false negatives. However, one then has to suffer from false positives. When $k_{min} = \infty$, any support in $\{\mathcal{S}_{omp}^k\}_{k=1}^{kmax}$ has to suffer from false negatives and all supports $\mathcal{S}_{omp}^k$ for $k > k_0 - 1$ has to suffer from false positives also. Note that the matrix and SNR conditions required for exact support recovery in Lemma 1 automatically implies that $k_{min} = k_0$. We formulate the proposed RRT scheme assuming that $k_{min} = k_0$.

3.2. Behaviour of $RR(k_0)$

Next we consider the behaviour of residual ratio statistic at the k_0 iteration, i.e., $RR(k_0) = \|\mathbf{r}^{k_0}\|_2/\|\mathbf{r}^{k_0-1}\|_2$ under the assumption that $\|\mathbf{w}\|_2 \leq \epsilon_{omp}$ and $\delta_{k_0+1} < 1/\sqrt{k_0+1}$ which ensures $k_{min} = k_0$ and $\mathcal{S}_{omp}^k \subseteq \mathcal{S}$ for all $k \leq k_0$. Since $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} \in span(\mathbf{X}_{\mathcal{S}})$, $(\mathbf{I}_n - \mathbf{P}_k)\mathbf{X}\boldsymbol{\beta} \neq \mathbf{0}_n$ if $\mathcal{S} \not\subseteq \mathcal{S}_{omp}^k$ and $(\mathbf{I}_n - \mathbf{P}_k)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}_n$ if $\mathcal{S} \subseteq \mathcal{S}_{omp}^k$. This along with the monotonicity of $\mathcal{S}_{omp}^k$ implies the following. $(\mathbf{I}_n - \mathbf{P}_k)\mathbf{X}\boldsymbol{\beta} \neq \mathbf{0}_n$ for $k < k_{min} = k_0$ and $(\mathbf{I}_n - \mathbf{P}_k)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}_n$ for $k \geq k_{min} = k_0$. Thus $\mathbf{r}^k = (\mathbf{I}_n - \mathbf{P}_k)\mathbf{y} = (\mathbf{I}_n - \mathbf{P}_k)\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} + (\mathbf{I}_n - \mathbf{P}_k)\mathbf{w}$ for $k < k_{min} = k_0$, whereas, $\mathbf{r}^k = (\mathbf{I}_n - \mathbf{P}_k)\mathbf{w}$ for $k \geq k_{min} = k_0$. Consequently, at $k = k_0$, the numerator $\|\mathbf{r}^{k_0}\|_2$ of $RR(k_0)$ contains contribution only from the noise term $\|(\mathbf{I}_n - \mathbf{P}_{k_0})\mathbf{w}\|_2$, whereas, the denominator $\|\mathbf{r}^{k_0-1}\|_2$ in $RR(k_0)$ contain contributions from both the signal term i.e., $(\mathbf{I}_n - \mathbf{P}_k)\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}$ and the noise term $(\mathbf{I}_n - \mathbf{P}_k)\mathbf{w}$. This behaviour of $RR(k_0)$ along with the fact that $\|\mathbf{w}\|_2 \xrightarrow{P} 0$ as $\sigma^2 \to 0$ implies the following theorem.

Theorem 1. *Assume that the matrix $\mathbf{X}$ satisfies the RIC constraint $\delta_{k_0+1} < 1/\sqrt{k_0+1}$ and $k_{max} > k_0$. Then*
a). $RR(k_{min}) \xrightarrow{P} 0$ as $\sigma^2 \to 0$.
b). $\lim_{\sigma^2\to 0} \mathbb{P}(k_{min} = k_0) = 1$.

Algorithm 2 Residual ratio thresholding

Input: Observation $\mathbf{y}$, matrix $\mathbf{X}$
Step 1: Run k_{max} iterations of OMP.
Step 2: Compute $RR(k)$ for $k = 1, \ldots, k_{max}$.
Step 3: Estimate $k_{RRT} = \max\{k : RR(k) \leq \Gamma^{\alpha}_{RRT}(k)\}$
Output: Support estimate $\hat{S} = \mathcal{S}^{k_{RRT}}_{omp}$. Vector estimate $\hat{\boldsymbol{\beta}}(\mathcal{S}^{k_{RRT}}_{omp}) = \mathbf{X}^{\dagger}_{\mathcal{S}^{k_{RRT}}_{omp}}\mathbf{y}$, $\hat{\boldsymbol{\beta}}(\{1, \ldots, p\}/\mathcal{S}^{k_{RRT}}_{omp}) = \mathbf{0}_{p-k_{RRT}}$.

3.3. Behaviour of $RR(k)$ for $k > k_{min}$

Next we discuss the behaviour of $RR(k)$ for $k > k_{min}$. By the definition of k_{min} we have $\mathcal{S} \subseteq \mathcal{S}^{k}_{omp}$ which implies that $\mathbf{r}^k = (\mathbf{I}_n - \mathbf{P}_k)\mathbf{w}$ for $k \geq k_{min}$. The absence of signal terms in numerator and the denominator of $RR(k) = \frac{\|(\mathbf{I}_n - \mathbf{P}_k)\mathbf{w}\|_2}{\|(\mathbf{I}_n - \mathbf{P}_{k-1})\mathbf{w}\|_2}$ for $k > k_{min}$ implies that even when $\|\mathbf{w}\|_2 \to 0$ or $\sigma^2 \to 0$, $RR(k)$ for $k > k_{min}$ does not converge to zero. This behaviour of $RR(k)$ for $k > k_{min}$ is captured in Theorem 2 where we provide explicit σ^2 or SNR independent lower bounds on $RR(k)$ for $k > k_{min}$.

Theorem 2. *Let $F_{a,b}(x)$ denotes the cumulative distribution function of a $\mathbb{B}(a, b)$ random variable. Then $\forall \sigma^2 > 0$,* $\Gamma^{\alpha}_{RRT}(k) = \sqrt{F^{-1}_{\frac{n-k}{2},0.5}\left(\frac{\alpha}{k_{max}(p-k+1)}\right)}$ *satisfies*

$$\mathbb{P}(RR(k) > \Gamma^{\alpha}_{RRT}(k), \forall k > k_{min}) \geq 1 - \alpha. \quad (3)$$

Theorem 2 states that the residual ratio statistic $RR(k)$ for $k > k_{min}$ is lower bounded by the deterministic sequence $\{\Gamma^{\alpha}_{RRT}(k)\}^{k_{max}}_{k=k_{min}+1}$ with a high probability (for small values of α). Please note that k_{min} is itself a R.V. Note that the sequence $\Gamma^{\alpha}_{RRT}(k)$ is dependent only on the matrix dimensions n and p. Further, Theorem 2 does not make any assumptions on the noise variance σ^2 or the design matrix $\mathbf{X}$. Theorem 2 is extremely non trivial considering the fact that the support estimate sequence $\{\mathcal{S}^{k}_{omp}\}^{k_{max}}_{k=1}$ produced by OMP is adaptive and data dependent.

Lemma 2. *The following important properties of $\Gamma^{\alpha}_{RRT}(k)$ are direct consequences of the monotonicity of CDF and the fact that a Beta R.V take values only in $[0, 1]$.*
1). $\Gamma^{\alpha}_{RRT}(k)$ is defined only in the interval $\alpha \in [0, k_{max}(p-k+1)]$.
2). $0 \leq \Gamma^{\alpha}_{RRT}(k) \leq 1$.
3). $\Gamma^{\alpha}_{RRT}(k)$ is a monotonically increasing function of α.
4). $\Gamma^{\alpha}_{RRT}(k) = 0$ when $\alpha = 0$ and $\Gamma^{\alpha}_{RRT}(k) = 1$ when $\alpha = k_{max}(p-k+1)$.

3.4. Residual ratio thresholding framework

From Theorem 1, it is clear that $\mathbb{P}(k_{min} = k_0)$ and $\mathbb{P}(\mathcal{S}^{omp}_{k_0} = \mathcal{S})$ increases with increasing SNR (or decreasing σ^2), whereas, $RR(k_{min})$ decreases to zero with increasing SNR. At the same time, for small values of α like $\alpha = 0.01$, $RR(k)$ for $k > k_{min}$ is lower bounded by $\Gamma^{\alpha}_{RRT}(k)$ with a very high probability at all SNR. Hence, finding the last index k such that $RR(k) \leq \Gamma^{\alpha}_{RRT}(k)$, i.e., $k_{RRT} = \max\{k : RR(k) \leq \Gamma^{\alpha}_{RRT}(k)\}$ gives k_0 and equivalently $\mathcal{S}^{k_0}_{omp} = \mathcal{S}$ with a probability increasing with increasing SNR. This motivates the proposed signal and noise statistics oblivious RRT algorithm presented in Algorithm 2.

Remark 1. An important aspect regarding the RRT in Algorithm 2 is the choice of k_{RRT} when the set $\{k : RR(k) \leq \Gamma^{\alpha}_{RRT}(k)\} = \phi$. This situation happens only at very low SNR. When $\{k : RR(k) \leq \Gamma^{\alpha}_{RRT}(k)\} = \phi$ for a given value of α, we increase the value of α to the smallest value $\alpha_{new} > \alpha$ such that $\{k : RR(k) \leq \Gamma^{\alpha_{new}}_{RRT}(k)\} \neq \phi$. Mathematically, we set $k_{RRT} = \max\{k : RR(k) < \Gamma^{\alpha_{new}}_{RRT}(k)\}$, where $\alpha_{new} = \min_{a>\alpha}\{a : \{k : RR(k) \leq \Gamma^{\alpha}_{RRT}(k)\} \neq \phi\}$. Since $\alpha = p\, k_{max}$ gives $\Gamma^{\alpha}_{RRT}(1) = 1$ and $RR(1) \leq 1$, a value of $\alpha_{new} \leq pk_{max}$ always exists. α_{new} can be easily computed by first pre-computing $\{\Gamma^{a}_{RRT}(k)\}^{k_{max}}_{k=1}$ for say 100 prefixed values of a in the interval $(\alpha, pk_{max}]$.

Remark 2. RRT requires performing k_{max} iterations of OMP. All the quantities required for RRT including $RR(k)$ and the final estimates can be computed while performing these k_{max} iterations itself. Consequently, RRT has complexity $O(k_{max}np)$. As we will see later, a good choice of k_{max} is $k_{max} = [0.5(n+1)]$ which results in a complexity order $O(n^2p)$. This complexity is approximately n/k_0 times higher than the $O(npk_0)$ complexity of OMP when k_0 or σ^2 are known *a priori*. This is the computational cost being paid for not knowing k_0 or σ^2 *a priori*. In contrast, L fold CV requires running $(1 - 1/L)n$ iterations of OMP L times resulting in a $O(L(1-1/L)n^2p) = O(Ln^2p)$ complexity, i.e., RRT is L times computationally less complex than CV.

Remark 3. RRT algorithm is developed only assuming that the support sequence generated by the sparse recovery algorithm is monotonically increasing. Apart from OMP, algorithms such as orthogonal least squares(Wen et al., 2017) and OMP with thresholding(Yang & de Hoog, 2015) also produce monotonic support sequences. RRT principle can be directly applied to operate these algorithms in a signal and noise statistics oblivious fashion.

4. Analytical Results for RRT

In this section we present support recovery guarantees for RRT and compare it with the results available for OMP with *a priori* knowledge of k_0 or σ^2. The first result in this section deals with the finite sample and finite SNR performance for RRT.

Theorem 3. *Let $k_{max} \geq k_0$ and suppose that the matrix $\mathbf{X}$ satisfies $\delta_{k_0+1} < \frac{1}{\sqrt{k_0+1}}$. Then RRT can recover the true support $\mathcal{S}$ with probability greater than $1 - 1/n - \alpha$ provided that $\epsilon_\sigma < \min(\epsilon_{omp}, \epsilon_{rrt})$, where*

$$\epsilon_{rrt} = \frac{\Gamma^{\alpha}_{RRT}(k_0)\sqrt{1-\delta_{k_0}}\boldsymbol{\beta}_{min}}{1+\Gamma^{\alpha}_{RRT}(k_0)}. \tag{4}$$

Theorem 3 implies that RRT can identify the support $\mathcal{S}$ at a higher SNR or lower noise level than that required by OMP with *a priori* knowledge of k_0 and σ^2. For small values of α like $\alpha = 0.01$, the probability of exact support recovery, i.e., $1 - \alpha - 1/n$ is similar to that of the $1 - 1/n$ probability of exact support recovery in Lemma 1. Also please note that the RRT framework does not impose any extra conditions on the design matrix $\mathbf{X}$. Consequently, the only appreciable difference between RRT and OMP with *a priori* knowledge of k_0 and σ^2 is in the extra SNR required by RRT which is quantified next using the metric $\epsilon_{extra} = \epsilon_{omp}/\epsilon_{rrt}$. Note that the larger the value of ϵ_{extra}, larger should be the SNR or equivalently smaller should be the noise level required for RRT to accomplish exact support recovery. Substituting the values of ϵ_{omp} and ϵ_{rrt} and using the bound $\delta_{k_0} \leq \delta_{k_0+1}$ gives

$$\epsilon_{extra} \leq \frac{1+\frac{1}{\Gamma^{\alpha}_{RRT}(k_0)}}{1+\frac{\sqrt{1-\delta^2_{k_0+1}}}{1-\sqrt{k_0+1}\delta_{k_0+1}}}. \tag{5}$$

Note that $\frac{\sqrt{1-\delta^2_{k_0+1}}}{1-\sqrt{k_0+1}\delta_{k_0+1}} = \left(\frac{1-\delta_{k_0+1}}{1-\sqrt{k_0+1}\delta_{k_0+1}}\right)\sqrt{\frac{1+\delta_{k_0+1}}{1-\delta_{k_0+1}}} \geq 1$. Consequently,

$$\epsilon_{extra} \leq 0.5\left(1+\frac{1}{\Gamma^{\alpha}_{RRT}(k_0)}\right). \tag{6}$$

Since $0 \leq \Gamma^{\alpha}_{RRT}(k_0) \leq 1$, it follows that $0.5\left(1+\frac{1}{\Gamma^{\alpha}_{RRT}(k_0)}\right)$ is always greater than or equal to one. However, ϵ_{extra} decreases with the increase in $\Gamma^{\alpha}_{RRT}(k_0)$. In particular, when $\Gamma^{\alpha}_{RRT}(k_0) = 1$, there is no extra SNR requirement.

Remark 4. RRT algorithm involves two hyper parameters *viz.* k_{max} and α. Exact support recovery using RRT requires only that $k_{max} \geq k_0$. However, k_0 is an unknown quantity. In our numerical simulations, we set $k_{max} = \min(p, [0.5(rank(\mathbf{X})+1)])$. This choice is motivated by the facts that $k_0 < [0.5(rank(\mathbf{X})+1)]$ is a necessary condition for exact support recovery using any sparse estimation algorithm(Elad, 2010) when $n < p$ and $\min(n,p)$ is the maximum possible number of iterations in OMP. Since evaluating $rank(\mathbf{X})$ requires extra computations, one can always use $rank(\mathbf{X}) \leq n$ to set $k_{max} = \min(p, [0.5(n+1)])$. Please note that this choice of k_{max} is independent of the operating SNR, design matrix and the vector to be estimated and the user is not required to tune this parameter. Hence, α is the only user specified hyper parameter in RRT algorithm.

4.1. Large sample behaviour of RRT

Next we discuss the behaviour of RRT as $n \to \infty$. From (6), it is clear that the extra SNR required for support recovery using RRT decreases with increasing $\Gamma^{\alpha}_{RRT}(k_0)$. However, by Lemma 2 increasing $\Gamma^{\alpha}_{RRT}(k_0)$ requires an increase in the value of α. However, increasing α decreases the probability of support recovery given by $1 - \alpha - 1/n$. In other words, one cannot have exact support recovery using RRT at lower SNR without increasing the probability of error in the process. An answer to this conundrum is available in the large sample regime where it is possible to achieve both $\alpha \approx 0$ and $\Gamma^{\alpha}_{RRT}(k_0) \approx 1$, i.e., no extra SNR requirement and no decrease in probability of support recovery. The following theorem states the conditions required for $\Gamma^{\alpha}_{RRT}(k_0) \approx 1$ for large values of n.

Theorem 4. *Define $k_{lim} = \lim_{n\to\infty} k_0/n$, $p_{lim} = \lim_{n\to\infty} \log(p)/n$ and $\alpha_{lim} = \lim_{n\to\infty} \log(\alpha)/n$. Let $k_{max} = \min(p, [0.5(n+1)])$. Then $\Gamma^{\alpha}_{RRT}(k_0) = \sqrt{F^{-1}_{\frac{n-k_0}{2},0.5}\left(\frac{\alpha}{k_{max}(p-k_0+1)}\right)}$ satisfies the following asymptotic limits.*

Case 1:-). $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = 1$, *whenever* $k_{lim} < 0.5$, $p_{lim} = 0$ *and* $\alpha_{lim} = 0$.

Case 2:-). $0 < \lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) < 1$ *if* $k_{lim} < 0.5$, $\alpha_{lim} = 0$ *and* $p_{lim} > 0$. *In particular,* $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = \exp(\frac{-p_{lim}}{1-k_{lim}})$.

Case 3:- $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = 0$ *if* $k_{lim} < 0.5$, $\alpha_{lim} = 0$ *and* $p_{lim} = \infty$.

Theorem 4 states that all choices of (n, p, k_0) satisfying $p_{lim} = 0$ and $k_{lim} < 0.5$ can result in $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = 1$ provided that the parameter α satisfies $\alpha_{lim} = 0$. Note that $\alpha_{lim} = 0$ for a wide variety of α including $\alpha =$ constant, $\alpha = 1/n^{\delta}$ for some $\delta > 0$, $\alpha = 1/\log(n)$ etc. It is interesting to see which (n, p, k_0) scenario gives $p_{lim} = 0$ and $k_{lim} < 0.5$. Note that exact recovery in $n < p$ scenario is possible only if $k_0 \leq [0.5(n+1)]$. Thus, the assumption $k_{lim} < 0.5$ will be satisfied in all interesting problem scenarios.

Regime 1:- $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = 1$ in low dimensional regression problems with p fixed and $n \to \infty$ or all $(n, p) \to (\infty, \infty)$ with $\lim_{n\to\infty} p/n \leq 1$.

Regime 2:- $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = 1$ in high dimensional case with p increases sub exponentially with n as $\exp(n^{\delta})$ for some $\delta < 1$ or p increases polynomially w.r.t n, i.e., $p = n^{\delta}$ for some $\delta > 1$. In both cases, $p_{lim} = \lim_{n\to\infty} \log(n^{\delta})/n = 0$ and $p_{lim} = \lim_{n\to\infty} \log(\exp(n^{\delta}))/n = 0$.

Regime 3:- $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = 1$ in the extreme high dimensional case where $(n, p, k_0) \to (\infty, \infty, \infty)$ satisfy-

ing $n \geq ck_0 \log(p)$ for some constant $c > 0$. Here $p_{lim} = \lim_{n\to\infty} \log(p)/n \leq \lim_{n\to\infty} \frac{1}{ck_0} = 0$ and $k_{lim} = \lim_{n\to\infty} 1/c\log(p) = 0$. Note that the sampling regime $n \approx 2k_0 \log(p)$ is the best known asymptotic guarantee available for OMP(Fletcher & Rangan, 2012).

Regime 4:- Consider a sampling regime where $(n, p) \to (\infty, \infty)$ such that k_0 is fixed and $n = ck_0 \log(p)$, i.e., p is exponentially increasing with n. Here $p_{lim} = 1/(ck_0)$ and $k_{\lim} = 0$. Consequently, $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = \exp\left(\frac{-1}{ck_0}\right) < 1$. A good example of this sampling regime is (Tropp & Gilbert, 2007) where it was shown that OMP can recover a (not every) particular k_0 dimensional signal from n random measurements (in noiseless case) when $n = ck_0 \log(p)$. Note that $c \leq 20$ for all k_0 and $c \approx 4$ for large k_0. Even if we assume that only $n = 4k_0 \log(p)$ measurements are sufficient for recovering a k_0 sparse signal, we have $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = \exp(-0.125) = 0.9512$ for $k_0 = 5$ (i.e., $\epsilon_{extra} \leq 1.0257$) and $\lim_{n\to\infty} \Gamma^{\alpha}_{RRT}(k_0) = \exp(-0.125) = 0.9753$ for $k_0 = 10$ (i.e.,$\epsilon_{extra} \leq 1.0127$).

Note that $\Gamma^{\alpha}_{RRT}(k_0) \to 1$ as $n \to \infty$ implies that $\epsilon_{extra} \to 1$ and $\min(\epsilon_{omp}, \epsilon_{rrt}) \to 1$. This asymptotic behaviour of $\Gamma^{\alpha}_{RRT}(k_0)$ and ϵ_{extra} imply the large sample consistency of RRT as stated in the following theorem.

Theorem 5. *Suppose that the sample size $n \to \infty$ such that the matrix $\mathbf{X}$ satisfies $\delta_{k_0+1} < \frac{1}{\sqrt{k_0+1}}$, $\epsilon_\sigma \leq \epsilon_{omp}$ and $p_{lim} = 0$. Then,*
a). OMP running k_0 iterations and OMP with SC $\|\mathbf{r}^k\|_2 \leq \epsilon_\sigma$ are large sample consistent, i.e.. $\lim_{n\to\infty} \mathbb{P}(\hat{\mathcal{S}} = \mathcal{S}) = 1$.
b). RRT with hyper parameter α satisfying $\lim_{n\to\infty} \alpha = 0$ and $\alpha_{lim} = 0$ is also large sample consistent.

Theorem 5 implies that at large sample sizes, RRT can accomplish exact support recovery under the same SNR and matrix conditions required by OMP with *a priori* knowledge of k_0 or σ^2. Theorem 5 has a very important corollary.

Remark 5. Theorem 1 implies that all choices of α satisfying $\alpha \to 0$ and $\alpha_{lim} = 0$ deliver similar performances as $n \to \infty$. Note that the range of adaptations satisfying $\alpha \to 0$ and $\alpha_{lim} = 0$ include $\alpha = 1/\log(n)$, $\alpha = 1/n^\delta$ for $\delta > 0$ etc. Since a very wide range of tuning parameters deliver similar results as $n \to \infty$, RRT is in fact asymptotically tuning free.

Remark 6. Based on the large sample analysis of RRT, one can make the following guidelines on the choice of α. When the sample size n is large, one can choose α as a function of n that satisfies both $\lim_{n\to\infty} \alpha = 0$ and $\alpha_{lim} = 0$. Also since the support recovery guarantees are of the form $1 - 1/n - \alpha$, it does not make sense to choose a value of α that decays to zero faster than $1/n$. Hence, it is preferable to choose values of α that decreases to zero slower than $1/n$ like $\alpha = 1/\log(n)$, $\alpha = 1/\sqrt{n}$ etc.

4.2. A high SNR operational interpretation of α

Having discussed the large sample behaviour of RRT, we next discuss the finite sample and high SNR behaviour of RRT. Define the events support recovery error $\mathcal{E} = \{\hat{\mathcal{S}} \neq \mathcal{S}\}$ and false positive $\mathcal{F} = card(\hat{\mathcal{S}}/\mathcal{S}) > 0$ and missed discovery or false negative $\mathcal{M} = card(\mathcal{S}/\hat{\mathcal{S}}) > 0$. The following theorem characterizes the likelihood of these events as SNR increases to infinity or $\sigma^2 \to 0$.

Theorem 6. *Let $k_{max} > k_0$ and the matrix $\mathbf{X}$ satisfies $\delta_{k_0+1} < 1/\sqrt{k_0+1}$. Then,*
a). $\lim_{\sigma^2\to 0} \mathbb{P}(\mathcal{M}) = 0$.
b). $\lim_{\sigma^2\to 0} \mathbb{P}(\mathcal{E}) = \lim_{\sigma^2\to 0} \mathbb{P}(\mathcal{F}) \leq \alpha$.

Theorem 6 states that when the matrix $\mathbf{X}$ allows for exact support recovery in the noiseless or low noise situation, RRT will not suffer from missed discoveries. Under such favourable conditions, α is a high SNR upper bound on both the probability of error and the probability of false positives. Please note that such explicit characterization of hyper parameters are not available for hyper parameters in Square root LASSO, RAT, LAT etc.

5. Numerical Simulations

In this section, we provide extensive numerical simulations comparing the performance of RRT with state of art sparse recovery techniques. In particular, we compare the performance of RRT with OMP with k_0 estimated using five fold CV and the least squares adaptive thresholding (LAT) proposed in (Wang et al., 2016). In synthetic data sets, we also compare RRT with OMP running exactly k_0 iterations and OMP with SC $\|\mathbf{r}^k\|_2 \leq \sigma\sqrt{n + 2\sqrt{n\log(n)}}$(Cai & Wang, 2011). These algorithms are denoted in Figures 1-4 by "CV", "LAT", "OMP1" and "OMP2" respectively. RRT1 and RRT2 represent RRT with parameter α set to $\alpha = 1/\log(n)$ and $\alpha = 1/\sqrt{n}$ respectively. By Theorem 5, RRT1 and RRT2 are large sample consistent.

5.1. Synthetic data sets

The synthetic data sets are generated as follows. We consider two models for the matrix $\mathbf{X}$. Model 1 sample each entry of the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ independently according to $\mathcal{N}(0, 1)$. Matrix $\mathbf{X}$ in Model 2 is formed by concatenating $\mathbf{I}_n$ with a $n \times n$ Hadamard matrix $\mathbf{H}_n$, i.e., $\mathbf{X} = [\mathbf{I}_n, \mathbf{H}_n]$. This matrix guarantee exact support recovery using OMP at high SNR once $k_0 < \frac{1+\sqrt{n}}{2}$(Elad, 2010). The columns of $\mathbf{X}$ in both models are normalised to have unit l_2-norm. Based on the choice of $\mathbf{X}$ and support $\mathcal{S}$, we conduct 4 experiments. Experiments 1-2 involve matrix of model 1 with (n, p) given

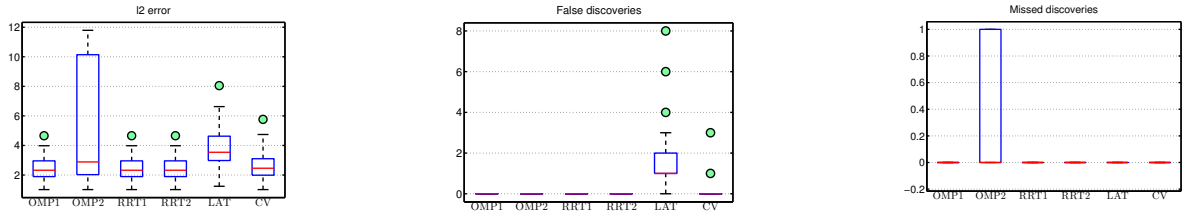

Figure 1. Experiment 1: Box plots of l_2 error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ (left), false positives (middle) and false negatives (right) .

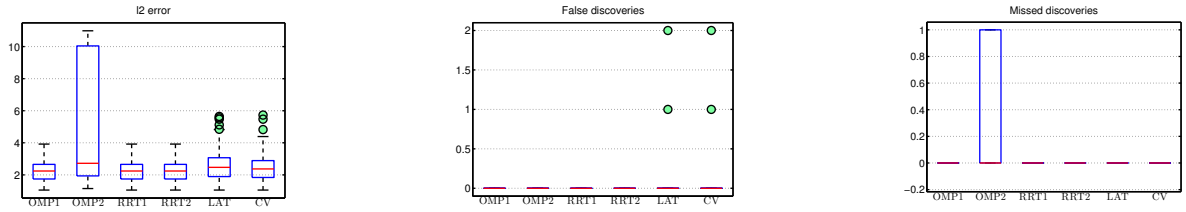

Figure 2. Experiment 2: Box plots of l_2 error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ (left), false positives (middle) and false negatives (right) .

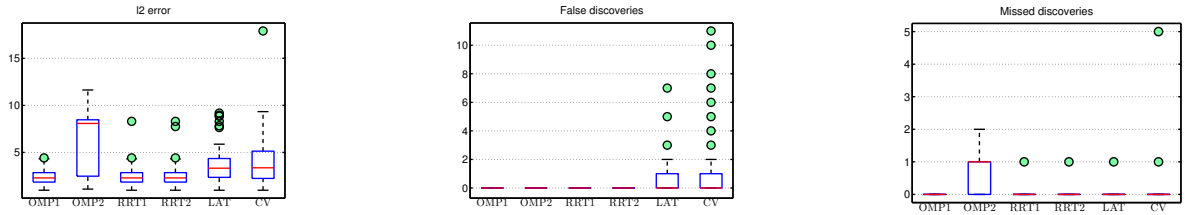

Figure 3. Experiment 3: Box plots of l_2 error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ (left), false positives (middle) and false negatives (right) .

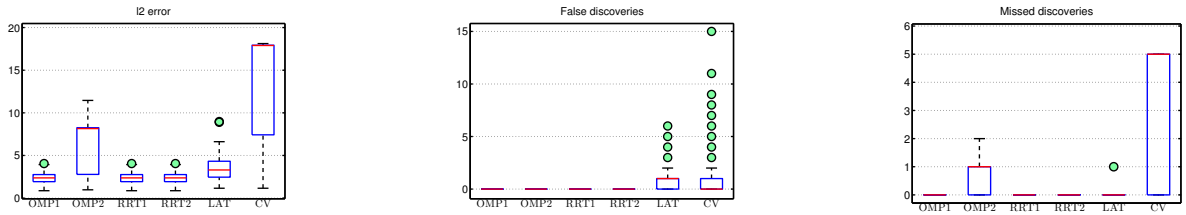

Figure 4. Experiment 4: Box plots of l_2 error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ (left), false positives (middle) and false negatives (right) .

Data Set	Outliers reported in literature	RRT	CV	LAT
Stack Loss $n = 21$ and $p = 4$ including intercept (Rousseeuw & Leroy, 2005)	1, 3, 4, 21 (Rousseeuw & Leroy, 2005)	1, 3, 4, 21	1, 3, 4, 21 plus 10 observations	4, 21
AR2000 $n = 60$ and $p = 3$ (Atkinson & Riani, 2012)	9, 21, 30, 31, 38, 47 (Atkinson & Riani, 2012)	9, 14, 21, 30 31, 38, 47, 50	9, 21, 30, 31, 38, 47 plus 41 observations	9, 14, 21 30, 31, 38 47, 50
Brain Body Weight $n = 27$ and $p = 1$ (Rousseeuw & Leroy, 2005)	1, 6, 14, 16, 17, 25 (Rousseeuw & Leroy, 2005)	1, 6, 16, 25	1, 6, 16, 25	1, 6, 16, 25
Stars $n_0 = 47$ and $p_0 = 1$ (Rousseeuw & Leroy, 2005)	11, 20, 30, 34 (Rousseeuw & Leroy, 2005)	11, 20, 30, 34	11, 20, 30, 34 plus 31 observations	11, 20, 30, 34

Table 1. Outliers detected by various algorithms. RRT with both $\alpha = 1/\log(n)$ and $\alpha = 1/\sqrt{n}$ delivered similar results. Existing results on Stack loss, Brain and Body weight and Stars data set are based on the combinatorially complex least median of squares (LMedS) algorithm. Existing results on AR2000 are based on extensive graphical analysis.

by $(200, 300)$ and $(200, 900)$ respectively with support $\mathcal{S}$ sampled randomly from the set $\{1, \ldots, p\}$. Experiment 3 and 4 involve matrix of model 2 with $(n = 128, p = 256)$. For experiment 3, support $\mathcal{S}$ is sampled randomly from the set $\{1, \ldots, p\}$, whereas, in experiment 4, support $\mathcal{S}$ is fixed at $\{1, 2, \ldots, k_0\}$. The noise $\mathbf{w}$ is sampled according to $\mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ with $\sigma^2 = 1$. The non zero entries of $\boldsymbol{\beta}$ are randomly assigned $\boldsymbol{\beta}_j = \pm 1$. Subsequently, these entries are scaled to achieve $SNR = \|\mathbf{X}\boldsymbol{\beta}\|_2^2/n = 3$. The number of non zero entries k_0 in all experiments are fixed at six. We compare the algorithms in terms of the l_2 error, the number of false positives and the number of false negatives produced in 100 runs of each experiment.

From the box plots given in Figures 1-4, it is clear that RRT with both values of α perform very similar to OMP1. They differ only in one run of experiment 3 where RRT1 and RRT2 suffer from a false negative. Further, RRT1 and RRT2 outperform CV and LAT in all the four experiments in terms of all the three metrics considered for evaluation. This is primarily because LAT and CV are more prone to make false positives, whereas RRT1 and RRT2 does not report any false positives. OMP2 consistently made false negatives which explains its poor performance in terms of l_2 error. We have observed that once the SNR is made slightly higher, OMP2 delivers a performance similar to OMP1. Also note that RRT with two significantly different choices of α *viz.* $\alpha = 1/\sqrt{n}$ and $\alpha = 1/\log(n)$ delivered similar performances. This observation is in agreement with the claim of asymptotic tuning freeness made in Remark 5. Similar trends are also visible in the simulation results presented in supplementary materials.

5.2. Outlier detection in real data sets

We next consider the application of sparse estimation techniques including RRT to identify outliers in low dimensional or full column rank (i.e., $n > p$) real life data sets, an approach first considered in (Mitra et al., 2010; 2013). Consider a robust regression model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w} + \mathbf{g}_{out}$ with usual interpretations for $\mathbf{X}$, $\boldsymbol{\beta}$ and $\mathbf{w}$. The extra term $\mathbf{g}_{out} \in \mathbb{R}^n$ represents the gross errors in the regression model that cannot be modelled using the distributional assumptions on $\mathbf{w}$. Outlier detection problem in linear regression refers to the identification of the support $\mathcal{S}_g = supp(\mathbf{g}_{out})$. Since $\mathbf{X}$ has full rank, one can always annihilate the signal component $\mathbf{X}\boldsymbol{\beta}$ by projecting onto a subspace orthogonal to $span(\mathbf{X})$. This will result in a simple linear regression model of the form given by

$$\tilde{\mathbf{y}} = (\mathbf{I}_n - \mathbf{X}\mathbf{X}^\dagger)\mathbf{y} = (\mathbf{I}_n - \mathbf{X}\mathbf{X}^\dagger)\mathbf{g}_{out} + (\mathbf{I}_n - \mathbf{X}\mathbf{X}^\dagger)\mathbf{w}, \tag{7}$$

i.e., identifying $\mathcal{S}_g$ in robust regression is equivalent to a sparse support identification problem in linear regression. Even though this is a regression problem with n observations and n variables, the design matrix $(\mathbf{I}_n - \mathbf{X}\mathbf{X}^\dagger)$ in (7) is rank deficient (i.e., $rank(\mathbf{I}_n - \mathbf{X}\mathbf{X}^\dagger) = n - rank(X) < n$). Hence, classical techniques based on LS are not useful for identifying $\mathcal{S}_g$. Since $card(\mathcal{S}_g)$ and variance of $\mathbf{w}$ are unknown, we only consider the application of RRT, OMP with CV and LAT in detecting $\mathcal{S}_g$. We consider four widely studied real life data sets and compare the outliers identified by these algorithms with the existing and widely replicated studies on these data sets. More details on these data sets are given in the supplementary materials. The outliers detected by the aforementioned algorithms and outliers reported in existing literature are tabulated in TABLE 1.

Among the four data sets considered, outliers detected by RRT and existing results are in consensus in two data sets *viz.* Stack loss and Stars data sets. In AR2000 data set, RRT identifies all the outliers. However, RRT also include observations 14 and 50 as outliers. These identifications can be potential false positives. In Brain and Body Weight data set, RRT agrees with the existing results in 4 observations. However, RRT misses two observations *viz.* 14 and 17 which are claimed to be outliers by existing results. LAT agrees with RRT in all data sets except the stack loss data set where it missed outlier indices 1 and 3. CV correctly identified all the outliers identified by other algorithms in all four data sets. However, it made lot of false positives in three data sets. To summarize, among all the three algorithms considered, RRT delivered an outlier detection performance which is the most similar to the results reported in literature.

6. Conclusions

This article proposed a novel signal and noise statistics independent sparse recovery technique based on OMP called residual ratio thresholding and derived finite and large sample guarantees for the same. Numerical simulations in real and synthetic data sets demonstrates a highly competitive performance of RRT when compared to OMP with *a priori* knowledge of signal and noise statistics. The RRT technique developed in this article can be used to operate sparse recovery techniques that produce a monotonic sequence of support estimates in a signal and noise statistics oblivious fashion. However, the support estimate sequence generated by algorithms like LASSO, DS, SP etc. are not monotonic in nature. Hence, extending the concept of RRT to operate sparse estimation techniques that produce non monotonic support sequence in a signal and noise statistics oblivious fashion is an interesting direction of future research.

References

Arlot, S., Celisse, A., et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4: 40–79, 2010.

Atkinson, A. and Riani, M. *Robust diagnostic regression analysis*. Springer Science & Business Media, 2012.

Bayati, M., Erdogdu, M. A., and Montanari, A. Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems*, pp. 944–952, 2013.

Belloni, A., Chernozhukov, V., and Wang, L. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

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

Candes, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pp. 2313–2351, 2007.

Chichignoud, M., Lederer, J., and Wainwright, M. J. A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *Journal of Machine Learning Research*, 17(231):1–20, 2016.

Dai, W. and Milenkovic, O. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, May 2009.

Dicker, L. H. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284, 2014.

Dicker, L. H. and Erdogdu, M. A. Maximum likelihood for variance estimation in high-dimensional linear models. In *Artificial Intelligence and Statistics*, pp. 159–167, 2016.

Elad, M. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

Fan, J., Guo, S., and Hao, N. Variance estimation using refitted cross-validation in ultra high dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

Fletcher, A. K. and Rangan, S. Orthogonal matching pursuit: A Brownian motion analysis. *IEEE Transactions on Signal Processing*, 60(3):1010–1021, March 2012.

Liu, C., Fang, Y., and Liu, J. Some new results about sufficient conditions for exact support recovery of sparse signals via orthogonal matching pursuit. *IEEE Transactions on Signal Processing*, 65(17):4511–4524, Sept 2017.

Mallat, S. G. and Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

Meinshausen, N. and Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Mitra, K., Veeraraghavan, A., and Chellappa, R. Robust regression using sparse learning for high dimensional parameter estimation problems. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3846–3849, March 2010.

Mitra, K., Veeraraghavan, A., and Chellappa, R. Analysis of sparse regularization based robust regression approaches. *IEEE Transactions on Signal Processing*, 61(5):1249–1257, March 2013.

Mousavi, A., Maleki, A., and Baraniuk, R. G. Parameterless optimal approximate message passing. *arXiv preprint arXiv:1311.0035*, 2013.

Pati, Y. C., Rezaiifar, R., and Krishnaprasad, P. S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pp. 40–44. IEEE, 1993.

Rousseeuw, P. J. and Leroy, A. M. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

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

Tropp, J. A. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006.

Tropp, J. A. and Gilbert, A. C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

Wang, J. Support recovery with orthogonal matching pursuit in the presence of noise. *IEEE Transactions on Information Theory*, 63(21):5868–5877, Nov 2015.

Wang, X., Dunson, D., and Leng, C. No penalty no tears: Least squares in high-dimensional linear models. In *International Conference on Machine Learning*, pp. 1814–1822, 2016.

Wen, J., Wang, J., and Zhang, Q. Nearly optimal bounds for orthogonal least squares. *IEEE Transactions on Signal Processing*, 65(20):5347–5356, 2017.

Yang, M. and de Hoog, F. Orthogonal matching pursuit with thresholding and its application in compressive sensing. *IEEE Transactions on Signal Processing*, 63(20):5479–5486, 2015.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.