Noisy and Missing Data Regression: Distribution-Oblivious Support Recovery

Yudong Chen YDCHEN@UTEXAS.EDU
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712

Constantine Caramanis CARAMANIS@MAIL.UTEXAS.EDU
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712

Abstract

Many models for sparse regression typically assume that the covariates are known completely, and without noise. Particularly in high-dimensional applications, this is often not the case. Worse yet, even estimating statistics of the noise (the noise covariance) can be a central challenge. In this paper we develop a simple variant of orthogonal matching pursuit (OMP) for precisely this setting. We show that without knowledge of the noise covariance, our algorithm recovers the support, and we provide matching lower bounds that show that our algorithm performs at the minimax optimal rate. While simple, this is the first algorithm that (provably) recovers support in a noise-distribution-oblivious manner. When knowledge of the noise-covariance is available, our algorithm matches the best-known ℓ^2-recovery bounds available. We show that these too are min-max optimal. Along the way, we also obtain improved performance guarantees for OMP for the standard sparse regression problem with Gaussian noise.

1. Introduction

Developing inference algorithms that are robust to corrupted or missing data is particularly important in the high-dimensional regime. There, not only can standard algorithms exhibit higher sensitivity to noisy/incomplete data, but often, the high-dimensionality may even preclude proper estimation of the statistics of the noise itself. Meanwhile, many standard algorithms for high-dimensional inference problems, including popular approaches such as ℓ^1-penalized regression, known as LASSO, are not equipped to deal with noisy or missing data.

This paper studies this precise problem, and focuses on sparsity recovery for linear regression, when the covariates are noisy or only partially known. Of particular interest is the noisy-covariates setting, where even the covariance of the noise is unknown. To the best of our knowledge, there are no algorithms that can recover sparsity without knowledge of the noise covariance. Indeed, as demonstrated in (Rosenbaum & Tsybakov, 2010), running standard algorithms like Lasso and the Dantzig selector without accounting for noise leads to unstable performance in support recovery. Moreover, we show via simulation that while provably optimal when the noise covariance is known, the two leading algorithms for regression with noisy covariates (Rosenbaum & Tsybakov, 2011; Loh & Wainwright, 2011) deteriorate rapidly with either an over- or under-estimate of the noise variance. However, in many practical scenarios, it is costly to obtain a reliable estimate of the noise covariance (Carroll, 2006), and in such situations there seems to be little that could be done. This paper delivers a surprising message. We show that a computationally much simpler algorithm, Orthogonal Matching Pursuit (OMP), succeeds in what we call *distribution oblivious* support recovery. Moreover, we prove that its performance is minimax optimal. In problems such as inverse covari-

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

ance estimation in graphical models, support recovery (finding the edges) is the central problem of interest. Meanwhile, that is precisely a setting where information about the noise may be difficult to estimate.

When the noise covariance is available, we show that a simple modification of OMP not only recovers the support, but also matches the ℓ^2-error guarantees of the more computationally demanding algorithms in (Rosenbaum & Tsybakov, 2011; Loh & Wainwright, 2011), and in particular, is optimal.

1.1. Related Work

The problem of sparse regression with noisy or missing covariates, in the high-dimensional regime, has recently attracted some attention, and several authors have considered this problem and made important contributions. Stadler and Buhlmann (Städler & Bühlmann, 2010) developed an EM algorithm to cope with missing data, but there does not seem to be a proof guaranteeing global convergence. Recent work has considered adapting existing approaches for sparse regression with good theoretical properties to handle noisy/missing data. The work in (Rosenbaum & Tsybakov, 2010; 2011) is among the first to obtain theoretical guarantees. They propose using a modified Dantzig selector (they called it the improved MU selector) as follows. Letting $y = X\beta + e$, and $Z = X + W$ denote the noisy version of the covariates (we define the setup precisely, below), the standard Dantzig selector would minimize $\|\beta\|_1$ subject to the condition $\|Z^\top(y - Z\beta)\|_\infty \leq \tau$. Instead, they solve $\|Z^\top(y - Z\beta) + \mathbb{E}[W^\top W]\beta\|_\infty \leq \mu\|\beta\|_1 + \tau$, thereby adjusting for the (expected) effect of the noise, W. Loh and Wainwright (Loh & Wainwright, 2011) pursue a related approach, where they modify Lasso rather than the Dantzig selector. Rather than minimize $\|Z\beta - y\|_2^2 + \lambda\|\beta\|_1$, they instead minimize a similarly adjusted objective: $\beta^\top(Z^\top Z - \mathbb{E}[W^\top W])\beta - 2\beta^\top Z^\top y + \|y\|_2^2 + \lambda\|\beta\|_1$. The modified Dantzig selector remains a linear program, and therefore can be solved by a host of existing techniques. The modified Lasso formulation becomes non convex. Interestingly, Loh and Wainwright show that the projected gradient descent algorithm finds a possibly local optimum that nevertheless has strong performance guarantees.

These methods obtain essentially equivalent ℓ^2-performance bounds, and recent work (Loh & Wainwright, 2012) shows they are minimax optimal. Significantly, they both rely on knowledge of $\mathbb{E}[W^\top W]$.[1] As our simulations demonstrate, this dependence seems critical: if the variance of the noise is either over- or under-estimated, the performance of the algorithms, even for support recovery, deteriorate considerably. The simple variant of the OMP algorithm we analyze requires no such knowledge for support recovery. Moreover, if $\mathbb{E}[W^\top W]$ is available, our algorithm has ℓ^2-performance matching that in (Rosenbaum & Tsybakov, 2011; Loh & Wainwright, 2011).

OMP has been studied extensively. Its performance in the clean covariate case has proven comparable to the computationally more demanding optimization-based methods, and it has been shown to be theoretically and empirically successful (e.g., (Tropp, 2004; Tropp & Gilbert, 2007; Davenport & Wakin, 2010)). OMP is an example of so-called greedy methods. Recently, there is a line of work that shows a particular class of forward-backward greedy algorithms is guaranteed to converge to the optimal solution of a convex program (Tewari et al., 2011; Jalali et al., 2011). Combining this work with guarantees on convex-optimization-based methods (e.g., (Loh & Wainwright, 2011)) could yield bounds on ℓ_2 recovery errors for noisy/missing data (this has not been done, but some variant of it seems straightforward). However, unlike our algorithm, this would still require knowing Σ_w and thus would not provide distribution-oblivious support recovery. Moreover, it is also worth pointing out that the forward-backward greedy methods required to guarantee support recovery under restricted strong convexity are different from, and more complicated than, OMP.

1.2. Contributions

We consider the problem of distribution-oblivious support recovery in high-dimensional sparse regression with noisy covariates. While the (unmodified) Lasso and Dantzig selector both require essential modification to work in the face of noise or missing data, we show a surprising result with simple and important consequences: standard (unmodified) OMP recovers the support, and the sample complexity and signal-to-noise ratio (the size of $\beta_{\min}$) required are optimal: we provide matching information theoretic lower bounds (see Theorem 4 for the precise statement). We then modify OMP so that if the noise covariance is available, our algorithm obtains ℓ^2-error bounds that match the best-known results ((Rosenbaum & Tsybakov, 2011; Loh & Wainwright, 2011)) and in particular, has matching lower bounds.

Specifically, the contributions of this paper are:

[1] The earlier work (Rosenbaum & Tsybakov, 2010) does not require that, but it does not guarantee support recovery, and its ℓ_2 error bounds are significantly weaker than the more recent work in (Rosenbaum & Tsybakov, 2011; Loh & Wainwright, 2011).

1. OMP and Support Recovery: We give conditions for when *standard OMP* guarantees exact support recovery in the missing and noisy covariate setting. Our results do not require knowledge of $\mathbb{E}[W^\top W]$ (the noise covariance) or ρ (the erasure probability). Other approaches require this knowledge, and our simulations indicate the dependence is real.

2. ℓ^2-bounds: Even if the noise covariance is known, standard OMP does not provide competitive ℓ^2-bounds. We design simple estimators based on either knowledge of the statistics of the covariate noise, or the covariate distribution. We provide finite sample performance guarantees that are as far as we know, the best available. In the supplemental section, we show we can also obtain such bounds using an Instrumental Variable correlated with the covariates. We are not aware of any rigorous finite sample results in this setting.

3. In simulations, the advantage of our algorithm seems more pronounced, in terms of both speed and statistical performance. Moreover, while we provide no analysis for the case of correlated columns of the covariate matrix, our simulations indicate that the impediment is in the analysis, as the results for our algorithm seem very promising.

4. Finally, as a corollary to our results above, setting the covariate-noise-level to zero, we obtain bounds on the performance of OMP in the standard setting, with additive Gaussian noise. Our bounds are better than bounds obtained by specializing deterministic results (e.g., ℓ^2-bounded noise as in (Donoho et al., 2006)) and ignoring Gaussianity; meanwhile, while similar to the results in (Cai & Wang, 2011), there seem to be gaps in their proof that we do not quite see how to fill.

As we mention above, while simulations indicate good performance even with correlated columns, the analysis we provide only guarantees bounds for the case of independent columns. In this respect, the convex optimization-based approaches in (Rosenbaum & Tsybakov, 2011; Loh & Wainwright, 2011) have an advantage. This advantage of optimization-based approaches over OMP is present even in the clean covariate case[2], and so does not seem to be a product of the missing/noisy covariates. Nevertheless, we believe it is a weakness in the analysis that must be addressed.

OMP-like methods, on the other hand, are significantly easier to implement, and computationally less demanding. Therefore, particularly for large-scale applications, understanding their performance is important, and they have a role to play even where optimization-based algorithms have proven successful. In this case, we have both the simplicity of OMP and the guarantee of distribution-oblivious recovery.

2. Problem Setup

We denote our unknown k-sparse regressor (or signal) in $\mathbb{R}^p$ as β^*. We obtain measurements $y_i \in \mathbb{R}$ according to the linear model

$$y_i = \langle \mathbf{x}_i, \beta^* \rangle + e_i, \quad i = 1, \ldots, n. \tag{1}$$

Here, $\mathbf{x}_i$ is a covariate vector of dimension p and $e_i \in \mathbb{R}$ is additive error. We are interested in the standard high-dimensional scaling, where $n = O(k \log p)$.

The standard setting assumes that each covariate vector $\mathbf{x}_i$ is known directly, and exactly. Instead, here we assume we only observe a vector $\mathbf{z}_i \in \mathbb{R}^p$ which is linked to $\mathbf{x}_i$ via some distribution that may be *unknown to the algorithm.* We focus on two cases:

1. Covariates with additive noise: We observe $\mathbf{z}_i = \mathbf{x}_i + \mathbf{w}_i$, where the entries of $\mathbf{w}_i \in \mathbb{R}^p$ (or $\mathbb{R}^k$) are independent of each other and everything else.

2. Covariates with missing data: We consider the case where the entries of $\mathbf{x}_i$ are observed independently with probability $1 - \rho$, and missing with probability ρ.

We consider the case where the covariate matrix X, the covariate noise W and the additive noise vector $\mathbf{e}$ are sub-Gaussian. We give the basic definitions here, as these are used throughout.

Definition 1. Sub-Gaussian Matrix: A zero-mean matrix V is called sub-Gaussian with parameter $(\frac{1}{n}\Sigma, \frac{1}{n}\sigma^2)$ if (a) Each row $\mathbf{v}_i^\top \in \mathbb{R}^p$ of V is sampled independently and has $\mathbb{E}\left[\mathbf{v}_i \mathbf{v}_i^\top\right] = \frac{1}{n}\Sigma$.[3] (b) For any unit vector $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{u}^\top \mathbf{v}_i$ is a sub-Gaussian random variable with parameter at most $\frac{1}{\sqrt{n}}\sigma$.

Definition 2. Sub-Gaussian Design Model: We assume X, W and $\mathbf{e}$ are sub-Gaussian with parameters $(\frac{1}{n}\Sigma_x, \frac{1}{n})$, $(\frac{1}{n}\Sigma_w, \frac{1}{n}\sigma_w^2)$ and $(\frac{1}{n}\sigma_e^2, \frac{1}{n}\sigma_e^2)$, respectively, and are independent of each other. For Section 3 we

[2] Standard deterministic analysis of OMP via RIP, e.g. (Davenport & Wakin, 2010), gives support recovery results with correlated columns, but with an extra $\sqrt{k}$ (where k is the sparsity) term, which cannot be removed completely (Rauhut, 2008)

[3] The $\frac{1}{n}$ factor is used to simplify subsequent notation; no generality is lost.

need independence across columns as well. We call this the *Independent* sub-Gaussian Model.

Our goal here consists of two tasks: 1) recovering the support of the unknown regressor β^*, and 2) finding a good estimate of β^* with small ℓ^2 error. These two tasks are typically considered simultaneously in the setting where X is observed without noise. They are quite different here. As mentioned above, we show that support recovery seems to be easier, and can be accomplished when the distribution linking the true covariate $\mathbf{x}_i$ and the observed covariate $\mathbf{z}_i$, is *unknown*.

3. Distribution-Oblivious Support Recovery via Orthogonal Matching Pursuit

In the case of additive noise, both the modified Dantzig selector and the modified Lasso can be viewed as trying to find the ℓ^1 constrained solution that satisfies the first order condition $(Z^\top Z - \Sigma_w)\beta - Z^\top y \approx 0$. Knowing Σ_w is necessary in order to construct the matrix $Z^\top Z - \Sigma_w$, which is an unbiased estimator of Σ_x; otherwise, solving the uncompensated condition $Z^\top Z\beta - Z^\top y \approx 0$ will consistently underestimate β^* even in the low-dimensional case. This seems unavoidable if the goal is to estimate the values of β^*.

What can we do when Σ_w is unknown – a setting we believe is particularly relevant in the high dimensional regime? For many applications (e.g., covariance estimation) support recovery alone is an important task. The main result in this paper is that this is possible, even when Σ_w is unknown.

To gain some intuition, first consider the simpler algorithm of One-Step Thresholding (OST) (Bajwa et al., 2010). OST computes the inner product between each column of Z and y, and finds the k columns with the largest inner product in magnitude (or those with inner product above some threshold.) This inner product does not involve the matrix $Z^\top Z$, and in particular, does not require compensation via $\mathbb{E}[W^\top W]$. Moreover, in expectation, the inner product equals the clean inner product. Thus under appropriate independence assumptions, the success of OST is determined by measure concentration rates.

This motivates us to use Orthogonal Matching Pursuit (OMP), which is more effective than the above simple algorithm, less sensitive to a dynamic range in nonzero entries of β^*, but still enjoys the distribution-oblivious property. We note again that empirically, our algorithm seems to succeed even with correlated columns – something we would not expect from OST.

Algorithm 1 support-OMP

```
Input: $Z$,$y$,$k$
Initialize $I = \phi$, $I^c = \{1, 2, \ldots, p\}$, $r = 0$.
for $j = 1 : k$ do
  Compute inner products $h_i = Z_i^\top r$, for $i \in I^c$.
  Let $i^* \leftarrow \arg\max_{i \in I^c} |h_i|$.
  Update support: $I \leftarrow I \cup \{i^*\}$.
  Update residual: $r \leftarrow y - Z_I(Z_I^\top Z_I)^{-1} Z_I^\top y$.
end for
Output: $I$.
```

3.1. Support Identification

Given a matrix Y and index set I, we use Y_I to denote the sub matrix with columns of Y indexed by I. We consider the following OMP algorithm, given in Algorithm 1. We call this supp-OMP to emphasize that its output is the support set, not $\hat{\beta}$. At each iteration, it computes the inner product between Z_i and the *residual* r_t instead of the original y, and picks the index with the largest magnitude.

We give some intuition on why one should expect supp-OMP would work without knowing Σ_w. Let I_t be the set of columns selected in the first t iterations of OMP; we assume the first t iterations succeed so $I_t \subset I^*$. Observe that OMP succeeds as long as the following two conditions are satisfied

1. $\langle Z_i, r_t \rangle = 0$, for all $i \in I_t$.
2. $\langle Z_i, r_t \rangle > \langle Z_j, r_t \rangle$, for all $i \in I^*/I_t$ and $j \in (I^*)^c$.

The first condition is satisfied automatically, due to the way we compute the residual $r_t = (I - \mathcal{P}_t)\, y \triangleq \left(I - Z_{I_t}\left(Z_{I_t}^\top Z_{I_t}\right)^{-1} Z_{I_t}^\top\right) y$; in particular, we orthogonalize y against Z_{I_t}, not X_{I_t}. Now consider the second condition. Observe that, for each $i \in I_t^c$, we have

$$\begin{aligned}\langle Z_i, r_t \rangle &= \langle Z_i, y \rangle - \langle Z_i, \mathcal{P}_t y \rangle \\ &= \langle Z_i, y \rangle - Z_i^\top Z_{I_t} \tilde{\beta}_t\end{aligned}$$

where $\tilde{\beta}_t = \left(Z_{I_t}^\top Z_{I_t}\right)^{-1} Z_{I_t}^\top y$. We call this $\tilde{\beta}$ because it is not the best estimate one could obtain, if the noise covariance were known. Indeed, there is an $\mathbb{E}[W^\top W]$ term embedded inside $\tilde{\beta}$, which, as discussed, is an underestimate of β^* as it is produced without compensating for the effect of $W^\top W$. The key insight and idea of the proof, is to show that despite the presence of this error, the dominating term is unbiased, so *the ordering of the inner products is unaffected.*

The next theorem gives the performance of supp-OMP for distribution-oblivious support recovery. In the sequel we use $\gtrsim$ to mean that we discard constants that do not depend on any scaling quantities, and by with high probability we mean with probability at least $1 - C_1 p^{-C_2}$, for positive absolute constants C_1, C_2.

Theorem 3. *Under the Independent sub-Gaussian Design model and Additive Noise model, supp-OMP identifies the correct support of β^* with high probability, provided*

$$n \gtrsim (1+\sigma_w^2)^2 k \log p,$$
$$|\beta_i^*| \geq 16 \left(\sigma_w \left\|\beta^*\right\|_2 + \sigma_e\right) \sqrt{\frac{\log p}{n}},$$

for all $i \in supp(\beta^*)$.

One notices two features of this result. Setting $\sigma_w = 0$ in Theorem 3, we obtain results that seem to be better than previous probabilistic guarantees for OMP with Gaussian noise and clean covariates. [4]

When $\sigma_w > 0$, the bound on $\beta_{\min}$ depends on $\|\beta^*\|_2$ – an SNR bound we typically do not see. This turns out to be fundamental. Indeed, we show that both the sample complexity and the SNR bounds are, in a precise sense, the best possible. In particular, this dependence on $\|\beta^*\|_2$ is unavoidable, and not an artifact of the algorithm or the analysis. Our next theorem characterizes the performance limit of *any* algorithm, regardless of its computational complexity. To this end, we consider the following minimax error

$$\mathcal{M}_q \triangleq \min_{\hat{\beta}} \max_{\beta^* \in T} \mathbb{E} \left\| \hat{\beta} - \beta^* \right\|_q,$$

where the minimization is over all measurable functions $\hat{\beta}$ of the observed data (Z, y), and the expectation is over the distribution of (Z, y). In other words, no algorithm can achieve error lower than $\mathcal{M}_q$ in expectation. T is the set of possible β^*'s;we consider $T \triangleq \{\beta : \|\beta\|_0 = k, \|\beta\|_2 = R, |\beta_i| \geq b_{\min}, \forall i \in \operatorname{supp}(\beta)\}$; that is, β^* is known to be k-sparse, have 2-norm equal to R, with its non-zero entries having magnitude at least $b_{\min}$. We focus on $q = 2$ and $q = 0$, corresponding to ℓ^2 error and support recovery error. Note that $\|\hat{\beta} - \beta^*\|_0 > 0$ implies failure in support recovery. We have the following lower-bound on $\mathcal{M}_0$.

Theorem 4. *Let $\sigma_z^2 = \sigma_x^2 + \sigma_w^2$. Under the independent Gaussian model where the covariance of X, W and e are isotropic, if $n \lesssim \left(\sigma_w^2 + \frac{\sigma_z^2 \sigma_e^2}{R^2}\right) k \log\left(\frac{p}{k}\right)$ or $b_{\min} \lesssim \sqrt{\left(\sigma_w^2 R^2 + \sigma_z^2 \sigma_e^2\right) \frac{\log(p/k)}{n}}$, then $\mathcal{M}_0 \geq 1$.*

Note the dependence on $R = \|\beta^*\|_2$.

3.2. Estimating the Non-Zero Coefficients

Once the support of β^* is identified, estimating its non-zero values β_I^* becomes a low-dimensional problem. Given the output I of supp-OMP, we consider the following generic estimator

$$\begin{aligned} \hat{\beta}_I &= \hat{\Sigma}^{-1} \hat{\gamma}, \\ \hat{\beta}_{I^c} &= 0. \end{aligned} \tag{2}$$

The estimator requires computing two matrices, $\hat{\Sigma}$ and $\hat{\gamma}$. If X is known, setting $\hat{\Sigma} = X_I^\top X_I$ and $\hat{\gamma} = X_I^\top y$ gives the standard least squares estimator. For our problem where only Z is observed, some knowledge of X or the nature of the corruption (W in the case of additive noise) is required in order to proceed. For the case of additive noise, we consider two models for *a priori* knowledge of X or of the noise:

1. Knowledge of Σ_w: in this case, we assume we either know or somehow can estimate the noise covariance on the true support, $\Sigma_w^{(I)} = \mathbb{E}\left[W_I^\top W_I\right]$. We then use $\hat{\Sigma} = Z_I^\top Z_I - \Sigma_w^{(I)}$ and $\hat{\gamma} = Z_I^\top y$.

2. Knowledge of Σ_x: we assume that we either know or somehow can estimate the covariance of the true covariates on the true support, $\Sigma_x^{(I)} = \mathbb{E}\left[X_I^\top X_I\right]$. We then use $\hat{\Sigma} = \Sigma_x^{(I)}$ and $\hat{\gamma} = Z_I^\top y$.

In both cases, only the covariance of the columns of W (or X) in the set I (which has size $k \ll p$) is required. Thus in the setting where estimation of the covariance is possible but costly, we have significantly reduced burden. Our second model is not as common as the previous one, although it seems equally plausible to have an estimate of $\mathbb{E}\left[X^\top X\right]$ as of $\mathbb{E}\left[W^\top W\right]$.

For the case of partially missing data, we assume we know the erasure probability, which is easy to estimate directly (cf. (Rosenbaum & Tsybakov, 2011)).

3. Knowledge of ρ: we compute $\hat{\Sigma} = (Z_I^\top Z_I) \odot M$ and $\hat{\gamma} = \frac{1}{(1-\rho)} Z_I^\top y$, where $M \in \mathbb{R}^{|I| \times |I|}$ satisfies $M_{ij} = \frac{1}{1-\rho}$ if $i = j$ or $\frac{1}{(1-\rho)^2}$ otherwise, and $\odot$ denotes element-wise product.

Theorem 5. *Under the Independent sub-Gaussian Design model and Additive Noise model, the output of estimator (2) satisfies:*

1. *(Knowledge of Σ_w):* $\left\|\hat{\beta} - \beta^*\right\|_2 \lesssim \left[\left(\sigma_w + \sigma_w^2\right) \|\beta^*\|_2 + \sigma_e \sqrt{1 + \sigma_w^2}\right] \sqrt{\frac{k \log p}{n}}.$

[4] The work in (Cai & Wang, 2011) obtains a similar condition on the non-zeros of β^*, however, the proof of their Theorem 8 applies the results of their Lemma 3 to bound $\|X^\top(I - X_I(X_I^\top X_I)^{-1}X_i^\top)\mathbf{e})\|_\infty$. As far as we can tell, however, Lemma 3 applies only to $\|X^\top \mathbf{e}\|_\infty$ thanks to independence, which need not hold for the case in question.

2. *(Knowledge of Σ_x):* $\left\|\hat{\beta}-\beta^*\right\|_2 \lesssim \left[(1+\sigma_w)\left\|\beta^*\right\|_2+\sigma_e\sqrt{1+\sigma_w^2}\right]\sqrt{\frac{k\log p}{n}}$.

As with support recovery, we now consider fundamental performance limits, and give lower bounds that show that our results are minimax optimal, by characterizing the performance limit of *any* algorithm, regardless of its computational complexity. Under the same setup as in Theorem 4, we have:

Theorem 6. *Let $\sigma_z^2 = \sigma_x^2+\sigma_w^2$, and suppose $8 \le k \le p/2$ and $n \lesssim k\log(p/k)$. Under the independent Gaussian model where the covariance of X, W and e are isotropic, the ℓ^2-error is bounded below: $M_2 \gtrsim \sqrt{(\sigma_w^2 R^2+\sigma_z^2\sigma_e^2)\frac{k}{n}\log\left(\frac{p}{k}\right)}$.*

Note again that the bounds we provide in Theorem 5 have the same dependence on $\|\beta^*\|_2$ as the lower bound in Theorem 6.

Discussion

We note that when σ_w is bounded above by a constant (strictly positive SNR) and $k = o(p)$ (sublinear sparsity), the conditions for support recovery (bounds on ℓ^2-error) in Theorem 3 and 4 (Theorem 5 and 6, respectively) match up to constant factors. This highlights the tightness of our analysis.

It is instructive to compare our algorithm and guarantees with existing work in (Rosenbaum & Tsybakov, 2011) and (Loh & Wainwright, 2011). There they do not provide guarantees for support recovery when Σ_w is not known. For ℓ^2 error under the sub-Gaussian design model, our condition and bound in Theorem 3 match those in (Loh & Wainwright, 2011), and is at least as good as (Rosenbaum & Tsybakov, 2011) applied to stochastic settings (the difference is $\|\beta^*\|_2/\|\beta^*\|_1$, which is always no more than 1). The lower-bound for ℓ^2-error in Theorem 6 is essentially proved in (Loh & Wainwright, 2012), but the lower bound for support recovery in Theorem 4 is new.

Finally, we note that the number of iterations in supp-OMP depends on the sparsity k. In practice, one can use other stopping rules, e.g., those based on cross-validation. We do not pursue this here. The results in (Rosenbaum & Tsybakov, 2011) require two tuning parameters, μ and τ, which ultimately depend on knowledge of the noise W and e. Results in (Loh & Wainwright, 2011) require knowing the parameter $B = \|\beta^*\|_1$. Generally speaking, some forms of parameters like k, μ, τ and B are inevitable for any algorithm, and none of them is strictly easier to determine than the others. Moreover, if one has a good choice of one of these parameters, one can often determine the other parameters by, e.g., cross-validation.

3.3. Missing Data

We provide guarantees analogous to Theorem 3 above.

Theorem 7. *Under the Independent sub-Gaussian Design model and missing data model, supp-OMP identifies the correct support of β^* provided*

$$n \gtrsim \frac{1}{(1-\rho)^4}k\log p,$$

$$|\beta_i^*| \ge \frac{16}{1-\rho}\left(\|\beta^*\|_2+\sigma_e\right)\sqrt{\frac{\log p}{n}},$$

for all $i \in supp(\beta^*)$. Moreover, the output of estimator (2) with knowledge of ρ satisfies

$$\left\|\hat{\beta}-\beta^*\right\|_2 \lesssim \left(\frac{1}{(1-\rho)^2}\left\|\beta^*\right\|_2+\frac{1}{1-\rho}\sigma_e\right)\sqrt{\frac{k\log p}{n}}.$$

Note that the condition on $|\beta_i|$ as well as the bounds for the ℓ^2 error again depend on $\|\beta^*\|_2$, corresponding to a natural and necessary SNR requirement analogous to that in Theorem 3. Characterization of fundamental performance limits similar to Theorem 4 can also be obtained, although it is harder to get sharp bounds than in the additive noise case (cf. (Loh & Wainwright, 2012)). We omit this here due to space constraints.

Again we compare with existing work in (Loh & Wainwright, 2011; Rosenbaum & Tsybakov, 2011). They do not provide distribution-oblivious support recovery. For ℓ^2 error, our conditions and bound match those in the most recent work in (Loh & Wainwright, 2012).

4. Proofs

The proofs of our results rely on careful use of concentration inequalities and information theoretic tools. We provide the key ideas here, by giving the proof of Theorems 3 and 4. We postpone the full details to the supplementary materials.

Proof of Theorem 3. The proof idea is as follows. We use induction, with the inductive assumption that the previous steps identify a subset I of the true support $I^* = \text{supp}(\beta^*)$. Let $I_r = I^* - I$ be the remaining true support that is yet to be identified. We need to prove that at the current step, supp-OMP picks an index in I_r, i.e., $\|h_{I_r}\|_\infty > |h_i|$ for all $i \in (I^*)^c$. We use a decoupling argument similar to (Tropp & Gilbert, 2007), showing that our supp-OMP identifies I^* if it identifies it in the same order as an oracle that runs supp-OMP only over I^*. Therefore we can assume I to be independent of X_i and W_i for all $i \in (I^*)^c$.

Define $\mathcal{P}_I \triangleq Z_I(Z_I^\top Z_I)^{-1}Z_I^\top$. By the triangle inequality, we have $\|h_{I_r}\|_\infty$ is lower-bounded by

$$\frac{1}{\sqrt{k-i}}\left(\left\|X_{I_r}^\top(I-\mathcal{P}_I)X_{I_r}\beta_{I_r}^*\right\|_2-\left\|W_{I_r}^\top(I-\mathcal{P}_I)X_{I_r}\beta_{I_r}^*\right\|_2 -\left\|Z_{I_r}^\top(I-\mathcal{P}_I)(W_I\beta_I-\mathbf{e})\right\|_2\right).$$

We would like to lower-bound the expression above. The main technical difficulty here is dealing with the randomness in set I. To this end, we use concentration to prove the uniform upper bound $\lambda_{\min}\left(X_{I_1^c}^\top(I-\mathcal{P}_{I_1})X_{I_1^c}\right) \geq \frac{1}{2}$ for all subsets $I_1 \subseteq I^*$ and $I_1^c \triangleq I^* - I_1$, which enables us to lower-bound the first term above. Similarly we upper-bound the next two terms by proving $\lambda_{\max}\left(W_{I_1^c}^\top(I-\mathcal{P}_{I_1})X_{I_1^c}\right) \leq \frac{1}{8}$. Combining pieces we obtain $|h_{I_r}|_\infty \geq \frac{1}{8\sqrt{k-i}}\|\beta_{I_r}^*\|_2$. Using similar ideas, we obtain the upper-bounds $|h_i| < \frac{1}{8\sqrt{k-i}}\|\beta_{I_r}^*\|_2$ for $i \in (I^*)^c$. This completes the proof of support recovery. Bounding ℓ^2 error is a low-dimensional problem, and the proof is given in the supplementary materials.

The proof of Theorem 7 follows from similar ideas.

Proof of Theorem 4. We use a standard argument to convert the problem of bounding the minimax error to a hypothesis testing problem. Let $P = \{\beta_1, \ldots, \beta_M\}$ be a (δ, p) packing set of the target set T, meaning $P \subseteq T$ and for all $\beta_j, \beta_l \in P$, $j \neq l$, we have $\|\beta_j - \beta_l\|_p \geq \delta$. Following (Yu, 1997; Yang & Barron, 1999; Raskutti et al., 2009), we have

$$\min_{\hat{\beta}} \max_{\beta^* \in T} \mathbb{E}\left\|\hat{\beta} - \beta^*\right\|_p \geq \frac{\delta}{2} \min_{\tilde{\beta}} \mathbb{P}\left(\tilde{\beta} \neq B\right), \qquad (3)$$

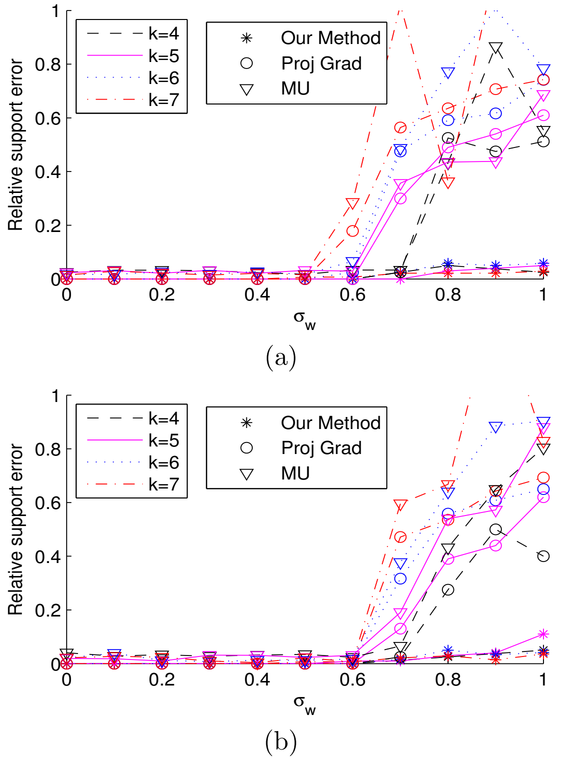
where $\tilde{\beta}$ is an estimator that takes values in P, and B is uniformly distributed over P. The probability on the R.H.S. can then be bounded by Fano's inequality:

$$\min_{\tilde{\beta}} \mathbb{P}\left(\tilde{\beta} \neq B\right) \geq 1 - \frac{I(y, Z; B) + \log 2}{\log M}.$$

The proof is completed by constructing an appropriate packing set P with sufficiently large cardinality M, followed by upper-bounding $I(y, Z; B)$ under the independent Gaussian Design model.

5. Numerical Simulations

In this section we provide numerical simulations that corroborate the theoretical results presented above, as well as shed further light on the performance of supp-OMP for noisy and missing data.

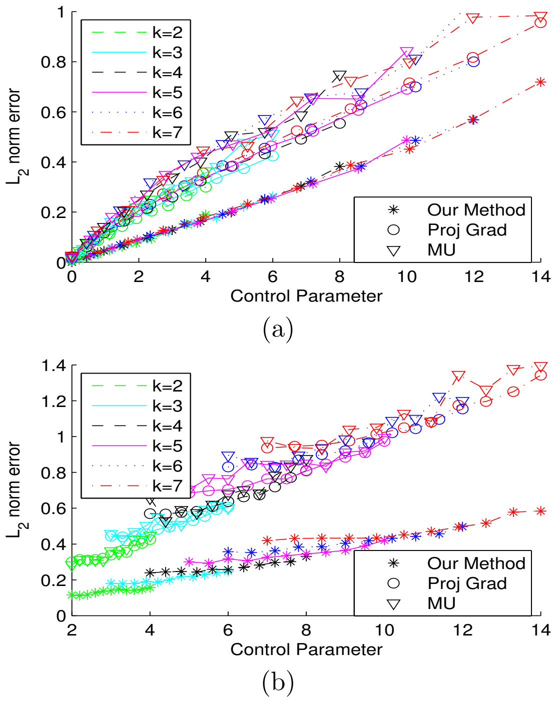


(a)

(b)

Figure 1. Support recovery errors when Σ_w is not precisely known. Our method, the projected gradient algorithm and the improved MU selector are shown. Figure (a) shows the results when we use $\Sigma_w = \Sigma_w^{\text{true}}/2$, and (b) shows the results for $\Sigma_w = \Sigma_w^{\text{true}} \times 2$. We note that our method (which does not require knowledge of Σ_w) has near perfect recovery throughout this range, whereas the performance of the other two methods deteriorates. Each point is an average over 20 trials.

The numerical results illustrate two main points. First, we show that the quality of support recovery using either the improved MU selector (or modified Dantzig selector) (Rosenbaum & Tsybakov, 2011) or the projected gradient algorithm based on modified Lasso (Loh & Wainwright, 2011), deteriorates significantly when Σ_w is not known accurately. That is, if those algorithms use either upper or lower bounds on Σ_w instead of the correct value, the results can be significantly distorted. In contrast, throughout the range of comparison, our supp-OMP algorithm is unaffected by the inaccuracy of Σ_w and recovers the support reliably. Next, we consider the bounds on ℓ^2-error, and show that the scaling promised in Theorem 5 is correct. In addition, we compare with the projected gradient algorithm and the improved MU selector, and demonstrate that in addition to faster running time, we seem to obtain better empirical results at all values of the sparsity parameter, and noise intensity/erasure probability. Beyond this, our simulations also show that our algorithm works well for correlated columns.



(a)

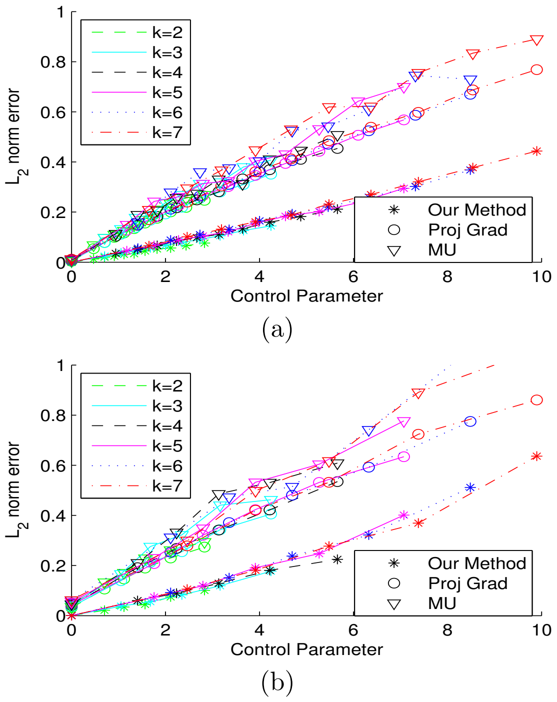
(b)

Figure 2. Comparison of the ℓ^2 recovery error of our method, the projected gradient algorithm and the improved MU selector under knowledge of (a) Σ_w, and (b) Σ_x. The error is plotted against the control parameter (a) $(\sigma_w + \sigma_w^2)k$, (b) $(1+\sigma_w)k$. Our method performs better in all cases. Each point is an average over 200 trials.

Additive Noise

For the additive noise case, we use the following settings: $p = 450, n = 400, \sigma_e = 0, \Sigma_x = I$, $k \in \{2, \ldots, 7\}$, and $\sigma_w \in [0, 1]$. The data X,W and e are generated from (isotropic) Gaussian distribution. We first consider support recovery, and study the robustness of the three methods to over- or under-estimating Σ_w. For low noise, the performance is largely unaffected; however, it quickly deteriorates as the noise level grows. The two graphs in Figure 1 show this deterioration; supp-OMP has excellent support recovery throughout this range.

Next we consider ℓ^2 recovery error. We run supp-OMP followed by the estimator 2 using the knowledge of Σ_w or Σ_x, and compare it with the other two methods using the same knowledge. Figure 2 plots the ℓ_2 errors for all three estimators. We note that although the estimator based on knowledge of Σ_x is not discussed in (Rosenbaum & Tsybakov, 2011) or (Loh & Wainwright, 2011), it is natural to consider the corresponding variants of their methods. One observes that supp-OMP outperforms the other two methods in all cases. We also want to point out that supp-OMP enjoys more favorable running time, as it has exactly the same running time as standard OMP.



(a)

(b)

Figure 3. Comparison of the ℓ^2 recovery error of our method, the projected gradient algorithm and the improved MU selector for missing data. The error is plotted against the control parameter $k\frac{\sqrt{\rho}}{(1-\rho)}$. (a) X with independent columns, and (b) X with correlated columns. Our results show that our method performs better in all cases in our simulations. Each point is an average over 50 trials.

Missing Data

We study the case with missing data with the following setting: $p = 750, n = 500, \sigma_e = 0, \Sigma_x = I$, $k \in \{2, \ldots, 7\}$, and $\rho \in [0, 0.5]$. The data X and e are generated from (isotropic) Gaussian distribution. The results are displayed in Figure 3, in which supp-OMP shows better performance over the other alternatives.

Finally, although we only consider X with independent columns in our theoretical analysis, simulations show our algorithm seems to work with correlated columns as well. Figure 3 (b) shows the results using covariance matrix of X: $(\Sigma_x)_{ij} = 1$ for $i = j$, and 0.2 for $i \neq j$. Again, supp-OMP dominates the other methods in terms of empirical performance.

In the supplementary materials section, we include further simulation results that detail the performance of our estimators in the low-dimensional setting.

References

Bajwa, W.U., Calderbank, R., and Jafarpour, S. Model selection: Two fundamental measures of coherence and their algorithmic significance. In *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, pp. 1568–1572. IEEE, 2010.

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

Carroll, R.J. *Measurement error in nonlinear models: a modern perspective*, volume 105. CRC Press, 2006.

Davenport, M.A. and Wakin, M.B. Analysis of orthogonal matching pursuit using the restricted isometry property. *Information Theory, IEEE Transactions on*, 56(9):4395–4401, 2010.

Donoho, D.L., Elad, M., and Temlyakov, V.N. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006.

Jalali, A., Johnson, C., and Ravikumar, P. On learning discrete graphical models using greedy methods. *arXiv preprint arXiv:1107.3258*, 2011.

Loh, P.L. and Wainwright, M.J. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Arxiv preprint arXiv:1109.3714*, 2011.

Loh, P.L. and Wainwright, M.J. Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pp. 2601–2605. IEEE, 2012.

Raskutti, G., Wainwright, M.J., and Yu, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *Arxiv preprint arXiv:0910.2042*, 2009.

Rauhut, H. On the impossibility of uniform sparse reconstruction using greedy methods. *Sampling Theory in Signal and Image Processing*, 7(2):197, 2008.

Rosenbaum, M. and Tsybakov, A.B. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

Rosenbaum, M. and Tsybakov, A.B. Improved matrix uncertainty selector. *arXiv preprint arXiv:1112.4413*, 2011.

Städler, N. and Bühlmann, P. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, pp. 1–17, 2010.

Tewari, A., Ravikumar, P., and Dhillon, I.S. Greedy algorithms for structurally constrained high dimensional problems. In *In Advances in Neural Information Processing Systems (NIPS) 24*, 2011.

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

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

Yang, Y. and Barron, A. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Yu, B. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pp. 423–435, 1997.