Submodular meets Spectral: Greedy Algorithms for Subset Selection, Sparse Approximation and Dictionary Selection

Abhimanyu Das ABHIMAND@USC.EDU
David Kempe DKEMPE@USC.EDU
Department of Computer Science,
University of Southern California

Abstract

We study the problem of selecting a subset of k random variables from a large set, in order to obtain the best linear prediction of another variable of interest. This problem can be viewed in the context of both feature selection and sparse approximation. We analyze the performance of widely used greedy heuristics, using insights from the maximization of submodular functions and spectral analysis. We introduce the *submodularity ratio* as a key quantity to help understand why greedy algorithms perform well even when the variables are highly correlated. Using our techniques, we obtain the strongest known approximation guarantees for this problem, both in terms of the submodularity ratio and the smallest k-sparse eigenvalue of the covariance matrix. We also analyze greedy algorithms for the dictionary selection problem, and significantly improve the previously known guarantees. Our theoretical analysis is complemented by experiments on real-world and synthetic data sets; the experiments show that the submodularity ratio is a stronger predictor of the performance of greedy algorithms than other spectral parameters.

1. Introduction

We analyze algorithms for the following important *Subset Selection* problem: select a subset of k variables from a given set of observation variables which, taken together, "best" predict another variable of interest. This problem has a wide range of applications ranging from feature selection, sparse learning and dictionary selection in machine learning, to sparse approximation and compressed sensing in signal processing. From a machine learning perspective, the variables could be features or observable attributes of a phenomenon, and we wish to predict the phenomenon using only a small subset from the high-dimensional feature space. In signal processing, the variables could correspond to a collection of dictionary vectors, and the goal is to parsimoniously represent another (output) vector. For many practitioners, the prediction model of choice is linear regression, and the goal is to obtain a linear model using a small subset of variables, to minimize the mean square prediction error or, equivalently, maximize the squared multiple correlation R^2 (Johnson & Wichern, 2002).

Thus, we formulate the Subset Selection problem for regression as follows: Given the (normalized) covariances between n variables X_i (which can in principle be observed) and a variable Z (which is to be predicted), select a subset of $k \ll n$ of the variables X_i and a linear prediction function of Z from the selected X_i that maximizes the R^2 fit. (A formal definition is given in Section 2.) The covariances are usually obtained empirically from detailed past observations of the variable values.

The above formulation is known (Das & Kempe, 2008) to be equivalent to the problem of *sparse approximation* over dictionary vectors: the input consists of a dictionary of n feature vectors $\mathbf{x_i} \in \mathbb{R}^m$, along with a target vector $\mathbf{z} \in \mathbb{R}^m$, and the goal is to select at most k vectors whose linear combination best approximates $\mathbf{z}$. The pairwise covariances of the previous formulation are then exactly the inner products of the dictionary vectors.[1]

Our problem formulation appears somewhat similar to the problem of sparse recovery (Zhang, 2008; 2009; Zhao & Yu, 2006; Candès et al., 2005); however, note that in sparse recovery, it is generally assumed that the prediction vector is truly (almost) k-sparse, and the aim is to recover the exact coefficients of this truly sparse solution. For

Appearing in *Proceedings of the 28^{th} International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

[1] For this reason, the dimension m of the feature vectors only affects the problem indirectly, via the accuracy of the estimated covariance matrix.

many situations, the actual solution might indeed be dense; the reason for running subset selection is then to reduce cost and model complexity.

This problem is **NP**-hard (Natarajan, 1995), so no efficient algorithms are known to solve it optimally. Two approaches are frequently used for approximating such problems: greedy algorithms (Miller, 2002; Tropp, 2004; Gilbert et al., 2003; Zhang, 2008) and convex relaxation schemes (Tibshirani, 1996; Candès et al., 2005; Tropp, 2006). For our formulation, a disadvantage of convex relaxation techniques is that they do not provide explicit control over the target sparsity level k of the solution; additional effort is needed to tune the regularization parameter.

A simpler and more intuitive approach, widely used in practice for subset selection problems (for example, it is implemented in all commercial statistics packages), is to use greedy algorithms, which iteratively add or remove variables based on simple measures of fit with Z. Two of the most well-known and widely used greedy algorithms are the subject of our analysis: Forward Regression (Miller, 2002) and Orthogonal Matching Pursuit (OMP) (Tropp, 2004). (These algorithms are defined in Section 2).

Previous theoretical bounds on such greedy algorithms have been unable to explain why they perform well in practice for most subset selection problem instances. Most previous results for greedy subset selection algorithms (Gilbert et al., 2003; Tropp, 2004; Das & Kempe, 2008) have been based on coherence of the input data, i.e., the maximum correlation μ between any pair of variables. Small coherence is an extremely strong condition, and the bounds (which usually prove a $1 - O(\mu k)$ approximation) break down when the coherence is $\omega(1/k)$. On the other hand, most bounds for greedy and convex relaxation algorithms for sparse recovery are based on a weaker sparse-eigenvalue or Restricted Isometry Property (RIP) condition (Zhang, 2009; 2008; Lozano et al., 2009; Zhou, 2009; Candès et al., 2005). However, these results apply to a different objective: minimizing the difference between the actual and estimated coefficients of a sparse vector. Simply extending these results to the subset selection problem adds a dependence on the largest k-sparse eigenvalue and only leads to weak additive bounds.

More importantly, all the above results rely on spectral conditions that fail to explain an observation of many experiments (including ours in Section 5): greedy algorithms often perform very well, even for near-singular input matrices. Our results begin to explain these observations by proving that the performance of many greedy algorithms does not really depend on how singular the covariance matrix is, but rather on how far the R^2 measure deviates from submodularity on the given input. We formalize this intuition by defining a measure of "approximate submodularity" which we term *submodularity ratio*. We prove that whenever the submodularity ratio is bounded away from 0, the R^2 objective is "reasonably close" to submodular, and Forward Regression gives a constant-factor approximation. This significantly generalizes our previous result (2008), where we had identified a strong condition termed "absence of conditional suppressors" which ensures that the R^2 objective is actually submodular.

An analysis based on the submodularity ratio does relate with traditional spectral bounds, in that this ratio is always lower-bounded by the smallest k-sparse eigenvalue of the covariance matrix (though it can be much larger when the predictor variable is not badly aligned with the eigenspace of small eigenvalues). In particular, we also obtain multiplicative approximation guarantees for Forward Regression and Orthogonal Matching Pursuit, whenever the smallest k-sparse eigenvalue is bounded away from 0, significantly strengthening past known bounds on their performance.

An added benefit of our framework is that we obtain tighter theoretical performance bounds for greedy algorithms for dictionary selection (Krause & Cevher, 2010). In the *dictionary selection problem* (which is formally defined in Section 2), we are given s target vectors, and a candidate set V of feature vectors. The goal is to select a set $D \subset V$ of at most d feature vectors, which will serve as a *dictionary* in the following sense. For each of the target vectors, the best $k < d$ vectors from D will be selected and used to achieve a good R^2 fit; the goal is to maximize the average R^2 fit for all of these vectors. This problem of finding a dictionary of basis functions for sparse representation of signals has several applications in machine learning and signal processing. Krause and Cevher (2010) showed that greedy algorithms for dictionary selection perform well in many instances, and proved additive approximation bounds for two specific algorithms, $\mathrm{SDS}_{\mathrm{MA}}$ and $\mathrm{SDS}_{\mathrm{OMP}}$. Our approximate submodularity framework lets us obtain stronger multiplicative guarantees without much extra effort.

Our theoretical analysis is complemented by experiments comparing the performance of the greedy algorithms and a baseline convex-relaxation algorithm for subset selection on two real-world data sets and a synthetic data set. More importantly, we evaluate the submodularity ratio of these data sets and compare it with other spectral parameters: while the input covariance matrices are close to singular, the submodularity ratio actually turns out to be significantly larger. Thus, our theoretical results can begin to explain why, in many instances, greedy algorithms perform well in spite of the fact that the data may have high correlations. Our main contributions can be summarized as follows:

1. We introduce the notion of submodularity ratio, as a much more accurate predictor of the performance of greedy algorithms than previously used parameters.

2. We obtain the strongest known theoretical performance guarantees for greedy algorithms for subset selection. In particular, we show (in Section 3) that the Forward Regression and OMP algorithms are within a $1 - e^{-\gamma}$ factor and $1 - e^{-(\gamma \cdot \lambda_{\min})}$ factor of optimal, respectively (where the γ and λ terms are appropriate submodularity and sparse-eigenvalue parameters).
3. We obtain the strongest known theoretical guarantees for algorithms for dictionary selection, improving on the results of (Krause & Cevher, 2010). In particular, we show (in Section 4) that the $\mathrm{SDS_{MA}}$ algorithm is within a factor $\frac{\gamma}{\lambda_{\max}}(1 - \frac{1}{e})$ of optimal.

2. Preliminaries

The goal in subset selection is to estimate a *predictor variable* Z using linear regression on a small subset from the set of *observation variables* $V = \{X_1, \ldots, X_n\}$. We use $\mathrm{Var}(X_i)$ and $\mathrm{Cov}(X_i, X_j)$ to denote the variance and covariance of the random variables. By appropriate normalization, we can assume that all the random variables have mean 0 and variance 1. The matrix of covariances between the X_i and X_j is denoted by C, with entries $c_{i,j} = \mathrm{Cov}(X_i, X_j)$. Similarly, we use $\mathbf{b}$ to denote the covariances between Z and the X_i, with entries $b_i = \mathrm{Cov}(Z, X_i)$. Formally, the *Subset Selection* problem can now be stated as follows:

Definition 2.1 (Subset Selection) *Given pairwise covariances among all variables, as well as a parameter k, find a set $S \subset V$ of at most k variables X_i and a linear predictor $Z' = \sum_{i \in S} \alpha_i X_i$ of Z, maximizing the* squared multiple correlation $R^2_{Z,S} := \frac{\mathrm{Var}(Z) - \mathrm{E}\left[(Z - Z')^2\right]}{\mathrm{Var}(Z)}$ *(Diekhoff, 2002; Johnson & Wichern, 2002).*

R^2 is a widely used measure for the goodness of a statistical fit; it captures the fraction of the variance of Z explained by variables in S. Because we assumed Z to be normalized to have variance 1, it simplifies to $R^2_{Z,S} = 1 - \mathrm{E}\left[(Z - Z')^2\right]$.

For a set S, we use C_S to denote the submatrix of C with row and column set S, and $\mathbf{b}_S$ to denote the vector with only entries b_i for $i \in S$. For notational convenience, we frequently do not distinguish between the index set S and the variables $\{X_i \mid i \in S\}$. Given the subset S of variables used for prediction, the optimal regression coefficients are well known to be $\alpha_S = (\alpha_i)_{i \in S} = C_S^{-1} \cdot \mathbf{b}_S$ (see, e.g., (Johnson & Wichern, 2002)), and hence $R^2_{Z,S} = \mathbf{b}_S^T C_S^{-1} \mathbf{b}_S$. Thus, the subset selection problem can be phrased as follows: Given C, $\mathbf{b}$, and k, select a set S of at most k variables to maximize $R^2_{Z,S} = \mathbf{b}_S^T (C_S^{-1}) \mathbf{b}_S$.[2]

The dictionary selection problem generalizes subset selection by considering s predictor variables $Z_1, Z_2, \ldots, Z_s$. The goal is to select a dictionary D of d observation variables, to optimize the average R^2 fit for the Z_j using at most k vectors from D for each. Formally, the Dictionary Selection problem is defined as follows:

Definition 2.2 (Dictionary Selection) *Given all pairwise covariances among the Z_j and X_i, and parameters d and k, find a set D of at most d variables from $\{X_1, \ldots, X_n\}$ maximizing $F(D) = \sum_{j=1}^{s} \max_{S_j \subset D, |S_j| = k} R^2_{Z_j, S_j}$.*

Many of our results are phrased in terms of eigenvalues of the covariance matrix C and its submatrices. Since covariance matrices are positive semidefinite, their eigenvalues are real and non-negative (Johnson & Wichern, 2002). For any positive semidefinite $n \times n$ matrix A, we denote its eigenvalues by $\lambda_{\min}(A) = \lambda_1(A) \leq \lambda_2(A) \leq \ldots \leq \lambda_n(A) = \lambda_{\max}(A)$. We use $\lambda_{\min}(C, k) = \min_{S:|S|=k} \lambda_{\min}(C_S)$ to refer to the smallest eigenvalue of any $k \times k$ submatrix of C (i.e., the smallest k-sparse eigenvalue), and similarly $\lambda_{\max}(C, k) = \max_{S:|S|=k} \lambda_{\max}(C_S)$. [3] We also use $\kappa(C, k)$ to denote the largest condition number (the ratio of the largest and smallest eigenvalue) of any $k \times k$ submatrix of C. This quantity is strongly related to the Restricted Isometry Property in (Candès et al., 2005). We also use $\mu(C) = \max_{i \neq j} |c_{i,j}|$ to denote the *coherence*, i.e., the maximum absolute pairwise correlation between the X_i variables. We denote $\mathrm{Res}(Z, S) = Z - \sum_{i \in S} \alpha_i X_i$ as the *residual* (see (Diekhoff, 2002)) of Z, i.e., the part of the variable Z that is not correlated with the X_i for all $i \in S$.

2.1. Submodularity Ratio

We introduce the notion of submodularity ratio for a general set function, which captures "how close" to submodular the function is. We first define it for arbitrary set functions, and then show the specialization for the R^2 objective.

Definition 2.3 (Submodularity Ratio) *Let f be a non-negative set function. The* submodularity ratio *of f with respect to a set U and a parameter $k \geq 1$ is $\gamma_{U,k}(f) = \min_{L \subseteq U, S: |S| \leq k, S \cap L = \emptyset} \frac{\sum_{x \in S} f(L \cup \{x\}) - f(L)}{f(L \cup S) - f(L)}$. Thus, it captures how much more f can increase by adding any subset S of size k to L, compared to the combined benefits of adding its individual elements to L.*

If f is specifically the R^2 objective defined on the variables X_i, then we omit f and simply write $\gamma_{U,k} = \min_{L \subseteq U, S: |S| \leq k, S \cap L = \emptyset} \frac{\sum_{i \in S} (R^2_{Z, L \cup \{X_i\}} - R^2_{Z,L})}{R^2_{Z, S \cup L} - R^2_{Z,L}} =$

[2] We assume throughout that C_S is non-singular. For some of our results, an extension to singular matrices is possible using the Moore-Penrose pseudoinverse.

[3] Computing $\lambda_{\min}(C, k)$ is NP-hard. In the full version of this paper (2011), we describe how to efficiently approximate the values for some scenarios.

$\min_{L\subseteq U,S:|S|\leq k,S\cap L=\emptyset} \frac{(\mathbf{b}_S^L)^T\mathbf{b}_S^L}{(\mathbf{b}_S^L)^T(C_S^L)^{-1}\mathbf{b}_S^L}$, *where* C^L *and* $\mathbf{b}^L$ *are the normalized covariance matrix and covariance vector corresponding to the set* $\{\mathrm{Res}(X_1,L),\mathrm{Res}(X_2,L),\ldots,\mathrm{Res}(X_n,L)\}$.

It can be easily shown that f is submodular if and only if $\gamma_{U,k}\geq 1$, for all U and k. For the purpose of subset selection, it is significant that the submodularity ratio can be bounded in terms of the smallest sparse eigenvalue, as shown by the following non-trivial lemma. (the proof is available in the full version of this paper (2011)).

Lemma 2.4 $\gamma_{U,k}\geq\lambda_{\min}(C,k+|U|)\geq\lambda_{\min}(C)$.

For all our analysis in this paper, we will use $|U|=k$, and hence $\gamma_{U,k}\geq\lambda_{\min}(C,2k)$. Thus, the smallest $2k$-sparse eigenvalue is a lower bound on this submodularity ratio; as we show later, it is often a weak lower bound.

3. Algorithms Analysis

We now present theoretical performance bounds for Forward Regression and Orthogonal Matching Pursuit. We also analyze the Oblivious algorithm, which is one of the simplest greedy algorithms for subset selection. Throughout this section, we use $\mathrm{OPT}=\max_{S:|S|=k}R^2_{Z,S}$ to denote the optimum R^2 value achievable by any set of size k. All proofs that are omitted due to space constraints can be found in the full version of this paper on arXiv (2011).

3.1. Forward Regression

We first provide approximation bounds for Forward Regression, which is the standard algorithm used by many researchers in medical, social and economic domains.

Definition 3.1 (Forward Regression) *The* Forward Regression *(also called* Forward Selection*) algorithm for subset selection selects a set S of size k iteratively as follows: 1) Initialize $S_0=\emptyset$. 2) In each iteration $i+1$, select the variable X_m maximizing $R^2_{Z,S_i\cup\{X_m\}}$, and set $S_{i+1}=S_i\cup\{X_m\}$. 3) Output S_k.*

Our main result is the following theorem. (The first inequality of our theorem can be shown to hold even if we replace $R^2_{Z,S}$ with an arbitrary monotone set function.)

Theorem 3.2 *The set S^{FR} selected by Forward Regression has the following approximation guarantees:* $R^2_{Z,S^{FR}}\geq(1-e^{-\gamma_{S^{FR},k}})\cdot OPT\geq(1-e^{-\lambda_{\min}(C,2k)})\cdot OPT\geq(1-e^{-\lambda_{\min}(C,k)})\cdot\Theta((\frac{1}{2})^{1/\lambda_{\min}(C,k)})\cdot OPT$.

Before proving the theorem, we first begin with a general lemma that bounds the amount by which the R^2 value of a set and the sum of R^2 values of its elements can differ.

Lemma 3.3 $\frac{1}{\lambda_{\max}(C)}\sum_{i=1}^n R^2_{Z,X_i}\leq R^2_{Z,\{X_1,\ldots,X_n\}}\leq\frac{1}{\gamma_{\emptyset,n}}\sum_{i=1}^n R^2_{Z,X_i}\leq\frac{1}{\lambda_{\min}(C)}\sum_{i=1}^n R^2_{Z,X_i}$.

Proof. Let the eigenvalues of C^{-1} be $\lambda'_1\leq\lambda'_2\leq\ldots\leq\lambda'_n$, with corresponding orthonormal eigenvectors $\mathbf{e_1},\mathbf{e_2},\ldots,\mathbf{e_n}$. We write $\mathbf{b}$ in the basis $\{\mathbf{e_1},\mathbf{e_2},\ldots,\mathbf{e_n}\}$ as $\mathbf{b}=\sum_i\beta_i\mathbf{e_i}$. Then, $R^2_{Z,\{X_1,\ldots,X_n\}}=\mathbf{b}^TC^{-1}\mathbf{b}=\sum_i\beta_i^2\lambda'_i$. Because $\lambda'_1\leq\lambda'_i$ for all i, we get $\lambda'_1\sum_i\beta_i^2\leq R^2_{Z,\{X_1,\ldots,X_n\}}$, and $\sum_i\beta_i^2=\mathbf{b}^T\mathbf{b}=\sum_i R^2_{Z,X_i}$, because the length of the vector $\mathbf{b}$ is independent of the basis it is written in. Also, by definition of the submodularity ratio, $R^2_{Z,\{X_1,\ldots,X_n\}}\leq\frac{\sum_i\beta_i^2}{\gamma_{\emptyset,n}}$. Finally, because $\lambda'_1=\frac{1}{\lambda_{\max}(C)}$, and using Lemma 2.4, we obtain the result. ∎

The next lemma relates the optimal R^2 value using k elements to the optimal R^2 using $k'<k$ elements.

Lemma 3.4 *For each k, let $S^*_k\in\mathrm{argmax}_{|S|\leq k}R^2_{Z,S}$ be an optimal subset of at most k variables.*

Then, for any $k'=\Theta(k)$ such that $\frac{1}{\lambda_{\min}(C,k)}<k'<k$, we have that $R^2_{Z,S^*_{k'}}\geq R^2_{Z,S^*_k}\cdot\Theta((\frac{k'}{k})^{1/\lambda_{\min}(C,k)})$, for large enough k. In particular, $R^2_{Z,S^*_{k/2}}\geq R^2_{Z,S^*_k}\cdot\Theta((\frac{1}{2})^{1/\lambda_{\min}(C,k)})$, for large enough k.

Using the above lemmas, we now prove the main theorem.

Proof of Theorem 3.2. We begin by proving the first inequality. Let S^*_k be the variables in the optimum solution. Let S^G_i be the set of variables chosen by Forward Regression in the first i iterations, and $S_i=S^*_k\setminus S^G_i$. By monotonicity of R^2 and the fact that $S_i\cup S^G_i\supseteq S^*_k$, we have that $R^2_{Z,S_i\cup S^G_i}\geq\mathrm{OPT}$.

For each $X_j\in S_i$, let $X'_j=\mathrm{Res}(X_j,S^G_i)$ be the residual of X_j conditioned on S^G_i, and write $S'_i=\{X'_j\mid X_j\in S\}$.

We will show that at least one of the X'_i is a good candidate in iteration $i+1$ of Forward Regression. First, the joint contribution of S'_i must be fairly large: $R^2_{Z,S'_i}=R^2_{Z,\mathrm{Res}(S_i,S^G_i)}\geq\mathrm{OPT}-R^2_{Z,S^G_i}$. Using Definition 2.3, $\sum_{X'_j\in S'_i}R^2_{Z,X'_j}\geq\gamma_{S^G_i,|S_i|}\cdot R^2_{Z,S'_i}\geq\gamma_{S^{\mathrm{FR}},k}\cdot R^2_{Z,S'_i}$, since $S^G_i\subseteq S^{\mathrm{FR}}$ and $|S_i|\leq k$. Let ℓ maximize R^2_{Z,X'_ℓ}, i.e., $\ell=\mathrm{argmax}_{(j:X'_j\in S'_i)}R^2_{Z,X'_j}$. Then we get that $R^2_{Z,X'_\ell}\geq\frac{\gamma_{S^{\mathrm{FR}},k}}{|S'_i|}\cdot R^2_{Z,S'_i}\geq\frac{\gamma_{S^{\mathrm{FR}},k}}{k}\cdot R^2_{Z,S'_i}$.

Define $A(i)=R^2_{Z,S^G_i}-R^2_{Z,S^{i-1}_G}$ to be the gain obtained from the variable chosen by Forward Regression in iteration i. Then $R^2_{Z,S^{\mathrm{FR}}}=\sum_{i=1}^k A(i)$. Since the X'_ℓ above was a candidate to be chosen in iteration $i+1$, and Forward Regression chose a variable X_m such that $R^2_{Z,\mathrm{Res}(X_m,S^G_i)}\geq$

$R^2_{Z,\mathrm{Res}(X,S_i^G)}$ for all $X \notin S_i^G$, we obtain that

$$\begin{aligned} A(i+1) &\geq \frac{\gamma_{S^{\mathrm{FR}},k}}{k} \cdot R^2_{Z,S'_i} \geq \frac{\gamma_{S^{\mathrm{FR}},k}}{k}(\mathrm{OPT} - R^2_{Z,S_i^G}) \\ &\geq \frac{\gamma_{S^{\mathrm{FR}},k}}{k}(\mathrm{OPT} - \textstyle\sum_{j=1}^{i} A(j)). \end{aligned}$$

Since the above inequality holds for each iteration $i = 1, 2, \ldots, k$, a simple inductive proof establishes the bound $\mathrm{OPT} - \sum_{i=1}^{k} A(i) \leq \mathrm{OPT} \cdot (1 - \frac{\gamma_{S^{\mathrm{FR}},k}}{k})^k$. Hence,

$$\begin{aligned} R^2_{Z,S^{\mathrm{FR}}} &= \textstyle\sum_{i=1}^{k} A(i) \geq \mathrm{OPT} - \mathrm{OPT}(1 - \frac{\gamma_{S^{\mathrm{FR}},k}}{k})^k \\ &\geq \mathrm{OPT} \cdot (1 - e^{-\gamma_{S^{\mathrm{FR}},k}}). \end{aligned}$$

The second inequality follows directly from Lemma 2.4, and the fact that $|S^{\mathrm{FR}}| = k$.

By applying the above result after $k/2$ iterations, we obtain $R^2_{Z,S^G_{k/2}} \geq (1 - e^{-\lambda_{\min}(C,k)}) \cdot R^2_{Z,S^*_{k/2}}$. Now, using Lemma 3.4 and monotonicity of R^2, we get $R^2_{Z,S^G_k} \geq R^2_{Z,S^G_{k/2}} \geq (1 - e^{-\lambda_{\min}(C,k)}) \cdot \Theta((\frac{1}{2})^{1/\lambda_{\min}(C,k)}) \cdot R^2_{Z,S^*_k}$, proving the third inequality. ■

3.2. Orthogonal Matching Pursuit

The second greedy subset selection algorithm that we analyze is Orthogonal Matching Pursuit (OMP), frequently used in signal processing domains.

Definition 3.5 (Orthogonal Matching Pursuit (OMP)) *The* OMP *algorithm selects a set S of size k iteratively as follows: 1) Initialize $S_0 = \emptyset$. 2) In each iteration $i+1$, select the variable X_m maximizing $|\mathrm{Cov}(\mathrm{Res}(Z, S_i), X_m)|$, and set $S_{i+1} = S_i \cup \{X_m\}$. 3) Output S_k.*

By applying similar techniques as in the previous section, we can also obtain approximation bounds for OMP.

Theorem 3.6 *The set S^{OMP} selected by Orthogonal Matching Pursuit has the following approximation guarantees: $R^2_{Z,S^{OMP}} \geq (1 - e^{-(\gamma_{S^{OMP},k} \cdot \lambda_{\min}(C,2k))}) \cdot OPT \geq (1 - e^{-\lambda_{\min}(C,2k)^2}) \cdot OPT \geq (1 - e^{-\lambda_{\min}(C,k)^2}) \cdot \Theta((\frac{1}{2})^{1/\lambda_{\min}(C,k)}) \cdot OPT$.*

For analyzing OMP, we first obtain the following lemma that lower-bounds the variance of the residual of a variable. This lemma, along with an analysis similar to the proof of Theorem 3.2, is then used to prove Theorem 3.6.

Lemma 3.7 *Let A be the $(n+1) \times (n+1)$ covariance matrix of the normalized variables $Z, X_1, X_2, \ldots, X_n$. Then* $\mathrm{Var}(\mathrm{Res}(Z, \{X_1, \ldots, X_n\})) \geq \lambda_{\min}(A)$

3.3. Oblivious Algorithm

As a baseline, we also consider a greedy algorithm which completely ignores C and simply selects the k variables individually most correlated with Z.

Definition 3.8 (Oblivious) *The* Oblivious *algorithm for subset selection is as follows: Select the k variables X_i with the largest b_i values.*

Lemma 3.3 immediately implies the following simple bound for the Oblivious algorithm.

Theorem 3.9 *The set S^{OBL} selected by the Oblivious algorithm has the following approximation guarantees: $R^2_{Z,S^{OBL}} \geq \frac{\gamma_{\emptyset,k}}{\lambda_{\max}(C,k)} \cdot OPT \geq \frac{\lambda_{\min}(C,k)}{\lambda_{\max}(C,k)} \cdot OPT$.*

4. Dictionary Selection Bounds

To demonstrate the wider applicability of the approximate submodularity framework, we next obtain a tighter analysis for two greedy algorithms for the dictionary selection problem, introduced in (Krause & Cevher, 2010).

The $\mathrm{SDS_{MA}}$ algorithm generalizes the Oblivious greedy algorithm to the problem of dictionary selection. It replaces the $R^2_{Z_j,S_j}$ term in Definition 2.2 with its modular approximation $f(Z_j, S_j) = \sum_{i \in S_j} R^2_{Z_j,X_i}$. Thus, it greedily tries to maximize the function $\hat{F}(D) = \sum_{j=1}^{s} \max_{S_j \subset D, |S_j| = k} f(Z_j, S_j)$, over sets D of size at most d; the inner maximum can be computed efficiently using the Oblivious algorithm.

Definition 4.1 ($\mathrm{SDS_{MA}}$) *The $\mathrm{SDS_{MA}}$ algorithm selects a dictionary D of size d iteratively as follows: 1) Initialize $D_0 = \emptyset$. 2) In each iteration $i+1$, select the variable X_m maximizing $\hat{F}(D_i \cup \{X_m\})$, and set $D_{i+1} = D_i \cup \{X_m\}$. 3) Output D_d.*

Using Lemma 3.3, we can obtain the following multiplicative approximation guarantee for $\mathrm{SDS_{MA}}$:

Theorem 4.2 *Let D^{MA} be the dictionary selected by the $\mathrm{SDS_{MA}}$ algorithm, and D^* the optimum dictionary of size $|D| \leq d$, with respect to the objective $F(D)$ from Definition 2.2. Then, $F(D^{MA}) \geq \frac{\gamma_{\emptyset,k}}{\lambda_{\max}(C,k)}(1 - \frac{1}{e}) \cdot F(D^*) \geq \frac{\lambda_{\min}(C,k)}{\lambda_{\max}(C,k)}(1 - \frac{1}{e}) \cdot F(D^*)$.*

Proof. Let $\hat{D}$ be a dictionary of size d maximizing $\hat{F}(D)$. Because f is monotone and modular, $\hat{F}$ is a monotone, submodular function. Hence, using the submodularity results of (Nemhauser et al., 1978) and the optimality of $\hat{D}$ for $\hat{F}$, $\hat{F}(D^{\mathrm{MA}}) \geq \hat{F}(\hat{D})(1 - \frac{1}{e}) \geq \hat{F}(D^*)(1 - \frac{1}{e})$. Now, by applying Lemma 3.3 for each Z_j, it is easy to show that $\hat{F}(D^*) \geq \gamma_{\emptyset,k} \cdot F(D^*)$, and similarly $\hat{F}(D^{\mathrm{MA}}) \leq \lambda_{\max}(C,k) \cdot F(D^{\mathrm{MA}})$. Thus we get $F(D^{\mathrm{MA}}) \geq \frac{\gamma_{\emptyset,k}}{\lambda_{\max}(C,k)}(1 - \frac{1}{e})F(D^*)$.

The second part now follows from Lemma 2.4. ■

Note that these bounds significantly improve the previous

additive approximation guarantee from (Krause & Cevher, 2010): $F(D^{\text{MA}}) \geq (1 - \frac{1}{e})F(D^*) - (2 - \frac{1}{e})k \cdot \mu(C)$. In particular, when $\mu(C) > \Theta(1/k)$, i.e., even just one pair of variables has moderate correlation, the approximation guarantee in (Krause & Cevher, 2010) becomes trivial.

We also obtain a multiplicative approximation guarantee for the greedy SDS_{OMP} algorithm, which improves on the additive bound obtained by Krause and Cevher. However, due to space constraints, the analysis is relegated to the full version of this paper (2011).

5. Experiments

In this section, we evaluate Forward Regression (FR) and OMP empirically, on two real-world and one synthetic data set. We plot the performance of the two algorithms against an optimal solution (OPT) computed using exhaustive search, the Oblivious greedy algorithm (OBL), and the L1-regularization/Lasso (L1) algorithm (using the implementation of (Koh et al., 2008)). Along with the observed performance, we also plot the theoretical performance bound for Forward Regression using the submodularity ratio (FR_γ), and that using the smallest sparse eigenvalue (FR_λ), which we had derived in Theorem 3.2. Additionally, we also compute and plot the various spectral parameters from which lower bounds can be derived. Specifically, these are (1) the submodularity ratio: $\gamma_{S^{\text{FR}},k}$, where S^{FR} is the subset selected by Forward Regression, (2) the smallest sparse eigenvalues $\lambda_{\min}(C, k)$ and $\lambda_{\min}(C, 2k)$. (In some cases, computing $\lambda_{\min}(C, 2k)$ was not feasible.) (3) the sparse inverse condition number $\kappa(C, k)^{-1}$, and (4) the smallest eigenvalue $\lambda_{\min}(C) = \lambda_{\min}(C, n)$ of the entire covariance matrix.

The aim of our experiments is twofold: First, we wish to evaluate which among the submodular and spectral parameters are good predictors of the performance of greedy algorithms in practice. Second, we wish to highlight how the theoretical bounds for subset selection algorithms reflect on their actual performance. Our analytical results predict that Forward Regression should outperform OMP, which in turn outperforms Oblivious. For Lasso, it is not known whether strong multiplicative bounds, like the ones we proved for Forward Regression or OMP, can be obtained.

5.1. Data Sets

Because several of the spectral parameters (as well as the optimum solution) are **NP**-hard to compute, we restrict our experiments to data sets with $n \leq 30$ features, from which $k \leq 8$ are to be selected. We stress that the greedy algorithms themselves are very efficient.

Each data set contains $m > n$ samples, from which we compute the empirical covariance matrix (analogous to the Gram matrix in sparse approximation) between all observation variables and the predictor variable; we then normalize it to obtain C and $\mathbf{b}$. We evaluate the performance of all algorithms in terms of their R^2 fit; thus, we implicitly treat C and $\mathbf{b}$ as the ground truth, and also do not separate the data sets into training and test cases.

Our data sets are the *Boston Housing Data*, a data set of *World Bank Development Indicators*, and a synthetic data set generated from a distribution similar to the one used by Zhang (2008). The *Boston Housing Data* (available from the UCI Machine Learning Repository) is a small data set frequently used to evaluate ML algorithms. It comprises $n = 15$ features (such as crime rate, property tax rates, etc.) and $m = 516$ observations. Our goal is to predict housing prices from these features. The *World Bank Data* (available from `http://databank.worldbank.org`) contains an extensive list of socio-economic and health indicators of development, for many countries and over several years. We choose a subset of $n = 29$ indicators for the years 2005 and 2006, such that the values for all of the $m = 65$ countries are known for each indicator. (The data set does not contain all indicators for each country.) We choose to predict the average life expectancy for those countries.

We also generate a synthetic dataset from a known distribution similar to (Zhang, 2008): There are $n = 29$ features, and $m = 100$ data points are generated from a joint Gaussian distribution with moderately high correlations of 0.6. The target vector is obtained by generating coefficients uniformly from 0 to 10 along each dimension, and adding noise with variance $\sigma^2 = 0.1$. Notice that the target vector is not truly sparse. The plots we show are the average R^2 values for 20 independent runs of the experiment.

5.2. Results

We run the various subset selection algorithms for values of k from 2 through 8, and plot the corresponding R^2 values. Figures 1, 3 and 5 show the results for the three data sets.

The main insight is that on all data sets, greedy algorithms perform exceedingly well compared to OPT. FR performs optimally or near-optimally, and OMP is only slightly worse in some cases. They are closely followed by Lasso, and then OBL. Interestingly, the order of observed performance of the greedy algorithms matches that of the strength of the theoretical bounds we derived for them.

In terms of the theoretical bounds for Forward Regression, in all cases, we see that our bound based on the submodularity ratio (FR_γ) is much stronger than that based on spectral parameters (FR_λ), which highlights the strength of our submodularity-based analysis. While there is still a substantial gap between FR_γ and the observed performance of FR, Section 5.3 shows how this gap can be narrowed.

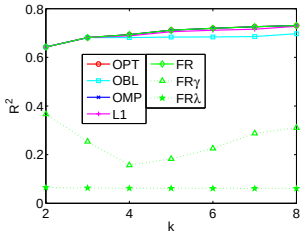

Figure 1: Boston Housing R^2

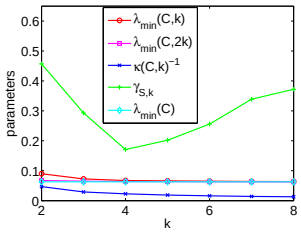

Figure 2: Boston Housing parameters

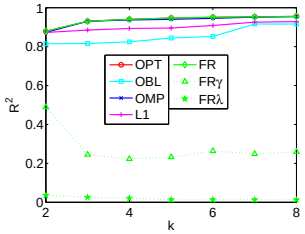

Figure 3: World Bank R^2

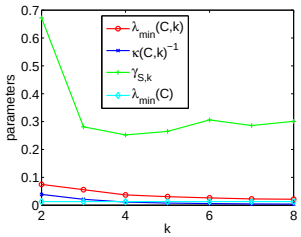

Figure 4: World Bank parameters

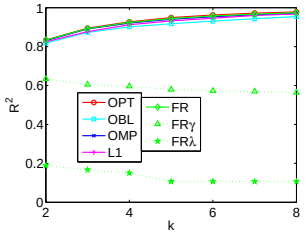

Figure 5: Synthetic Data R^2

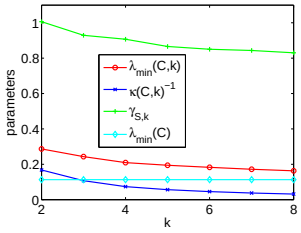

Figure 6: Synthetic Data parameters

Figures 2, 4 and 6 show the different spectral quantities for the data sets, for varying values of k. Both of the real-world data sets are nearly singular, as evidenced by the small $\lambda_{\min}(C)$ values. In fact, the near singularities manifest themselves for small values of k already; in particular, since $\lambda_{\min}(C, 2)$ is already small, we observe that there are pairs of highly correlated observations variables in the data sets. Thus, the bounds on approximation we would obtain by considering merely $\lambda_{\min}(C, k)$ or $\lambda_{\min}(C, 2k)$ would be quite weak (as observed by the plots for FR_λ in the previous set of graphs). Notice, however, that these parameters are still quite a bit stronger than the inverse condition number $\kappa(C, k)^{-1}$: this quantity — which is closely related to the RIP property that is frequently used in sparse recovery analysis — takes on much smaller values, and thus would lead to an even weaker bound.

The discrepancy between the small values of the eigenvalues and the good performance of all algorithms shows that eigenvalue-based bounds can frequently be loose. Significantly better bounds are obtained from the submodularity ratio $\gamma_{S^{\text{FR}},k}$ (as seen in the previous plots for FR_γ), which takes on values above 0.2, and significantly larger in some cases. While not entirely sufficient to explain the performance of the greedy algorithms, it shows that the near-singularities of C do not align unfavorably with $\mathbf{b}$, and thus do not provide an opportunity for strong supermodular behavior that adversely affects greedy algorithms.

The synthetic data set we generated is somewhat further from singular, with $\lambda_{\min}(C) \approx 0.11$. However, the same patterns persist: the simple eigenvalue based bounds, while somewhat larger for small k, still do not fully predict the performance of greedy algorithms, whereas the submodularity ratio here is close to 1 for all values of k. This shows that the near-singularities do not at all provide the possibility of strongly supermodular benefits of sets of variables. Indeed, the plot of R^2 values on the synthetic data is concave, an indicator of submodular behavior of the function.

The above observations suggest that bounds based on the submodularity ratio are better predictors of the performance of greedy algorithms, followed by bounds based on the sparse eigenvalues, and finally those based on the condition number or RIP property.

5.3. Narrowing the gap between theory and practice

Our theoretical bounds, though much stronger than previous results, still do not fully predict the observed near-optimal performance of Forward Regression and OMP on the real-world datasets. In particular, for Forward Regression, even though the submodularity ratio is less than 0.4 for most cases, implying a theoretical guarantee of roughly $1 - e^{-0.4} \approx 33\%$, the algorithm still achieves near-optimal performance. While gaps between worst-case bounds and practical performance are commonplace in algorithmic analysis, they also suggest that there is scope for further improving the analysis.

Indeed, a more careful analysis of the proof of Theorem 3.2 and our definition of the submodularity ratio reveals that we do not really need to calculate the submodularity ratio over all sets S of size k while analyzing the greedy steps of Forward Regression. We can ignore sets S whose submodularity ratio is low, but whose marginal contribution to the current R^2 is only a small fraction (say, at most ϵ). This is because the proof of Theorem 3.2 shows that if a particular greedy iteration uses such a set for its analysis, then the current solution must already be within a factor $\frac{1}{1+\epsilon}$ of the optimal solution. By carefully pruning such sets (using $\epsilon = 0.2$) while calculating the submodularity ratio, we see that the resulting values of $\gamma_{S^{\text{FR}},k}$ are much higher (more than 0.8), thus significantly reducing the gap between the theoretical bounds and experimental results. Table 1 shows the values of $\gamma_{S^{\text{FR}},k}$ obtained using this method.

The results suggest an interesting direction for future work: namely, to characterize for which sets the submodular behavior of R^2 really matters.

Table 1: Improved estimates for submodularity ratio

Data Set	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Boston	0.9	0.91	1.02	1.21	1.36	1.54	1.74
World Bank	0.8	0.81	0.81	0.81	0.94	1.19	1.40

6. Discussion and Concluding Remarks

In this paper, we analyze greedy algorithms using the notion of submodularity ratio, which captures how close to submodular an objective function (in our case the R^2 measure of statistical fit) is. Using submodular analysis, coupled with spectral techniques, we prove the strongest known approximation guarantees for commonly used greedy algorithms for subset selection and dictionary selection. Our bounds help explain why greedy algorithms perform well in practice even in the presence of strongly correlated data, and are substantiated by experiments on real-world and synthetic datasets. The experiments show that the submodularity ratio is a much stronger predictor of the performance of greedy algorithms than previously used spectral parameters. We believe that our techniques for analyzing greedy algorithms using a notion of "approximate submodularity" are not specific to subset selection and dictionary selection, and could also be used to analyze other problems in compressed sensing and sparse recovery.

Acknowledgments

We would like to thank Andreas Krause, Fei Sha and several anonymous referees for their helpful feedback.

References

Candès, E. J., Romberg, J., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59: 1207–1223, 2005.

Das, A. and Kempe, D. Algorithms for subset selection in linear regression. In *ACM Symposium on Theory of Computing*, 2008.

Das, A. and Kempe, D. Submodular meets Spectral: Greedy Algorithms for Sparse Approximation and Dictonary Selection, 2011. http://arxiv.org/abs/1102.3975.

Diekhoff, G. *Statistics for the Social and Behavioral Sciences.* Wm. C. Brown Publishers, 2002.

Gilbert, A., Muthukrishnan, S., and Strauss, M. Approximation of functions over redundant dictionaries using coherence. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2003.

Johnson, R. A. and Wichern, D. W. *Applied Multivariate Statistical Analysis.* Prentice Hall, 2002.

Koh, K., Kim, S., and Boyd, S. l1_ls: Simple Matlab Solver for l1-regularized Least Squares Problems, 2008. http://www.stanford.edu/ boyd/l1_ls.

Krause, A. and Cevher, V. Submodular dictionary selection for sparse representation. In *Proc. ICML*, 2010.

Lozano, A. C., Swirszcz, G., and Abe, N. Grouped orthogonal matching pursuit for variable selection and prediction. In *Proc. NIPS*, 2009.

Miller, A. *Subset Selection in Regression*. Chapman and Hall, second edition, 2002.

Natarajan, B. Sparse approximation solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

Nemhauser, G., Wolsey, L., and Fisher, M. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society*, 58:267–288, 1996.

Tropp, J. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Information Theory*, 50:2231–2242, 2004.

Tropp, J. Just relax: Convex programming methods for identifying sparse signals. *IEEE Trans. Information Theory*, 51:1030–1051, 2006.

Zhang, T. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Proc. NIPS*, 2008.

Zhang, T. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

Zhao, P. and Yu, B. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2451–2457, 2006.

Zhou, S. Thresholding procedures for high dimensional variable selection and statistical estimation. In *Proc. NIPS*, 2009.