Algorithms for Subset Selection in Linear Regression

Abhimanyu Das
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
abhimand@usc.edu

David Kempe*
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
dkempe@usc.edu

ABSTRACT

We study the problem of selecting a subset of k random variables to observe that will yield the best linear prediction of another variable of interest, given the pairwise correlations between the observation variables and the predictor variable. Under approximation preserving reductions, this problem is also equivalent to the "sparse approximation" problem of approximating signals concisely.

We propose and analyze exact and approximation algorithms for several special cases of practical interest. We give an FPTAS when the covariance matrix has constant bandwidth, and exact algorithms when the associated covariance graph, consisting of edges for pairs of variables with non-zero correlation, forms a tree or has a large (known) independent set. Furthermore, we give an exact algorithm when the variables can be embedded into a line such that the covariance decreases exponentially in the distance, and a constant-factor approximation when the variables have no "conditional suppressor variables".

Much of our reasoning is based on perturbation results for the R^2 multiple correlation measure, frequently used as a measure for "goodness-of-fit statistics". It lies at the core of our FPTAS, and also allows us to extend exact algorithms to approximation algorithms when the matrix "nearly" falls into one of the above classes. We also use perturbation analysis to prove approximation guarantees for the widely used "Forward Regression" heuristic when the observation variables are nearly independent.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

General Terms

Algorithms, Theory

*Supported in part by NSF CAREER award 0545855, and NSF grant DDDAS-TMRP 0540420

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
STOC'08, May 17–20, 2008, Victoria, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-047-0/08/05 ...$5.00.

1. INTRODUCTION

One of the most important algorithmic questions faced by data-driven sciences is to select the "most informative" subset from among a large set of observable attributes or parameters, to predict a particular quantity of interest. The setup is usually as follows: a large number of variables X_i can be (in principle) observed, and the researcher is interested in the value of a *predictor variable* Z. Due to time or cost constraints, it is not feasible to sample all the variables every time a prediction of Z is required. Frequently, offline, cost-intensive studies exist that reveal detailed information about the correlations between the X_i and Z. Based on this information, the goal is to select a (much smaller) subset of k variables X_i to predict Z in the future. This problem of selecting the k-subset of variables that "best" predicts Z is known as the *subset selection problem for regression* [21].

Natural applications of this problem abound. In medical or social studies, one often wants to predict risks or future behaviors (heart disease, failure in school, ...) in terms of observable quantities (blood pressure, parents' income, ...). The goal is to identify a small set of attributes for future tests. Similarly, in sensor networks, one often wants to sample a smaller number of sensors (in order to conserve energy), while obtaining accurate predictions of aggregate quantities.

In many of these scenarios, once a small set S of k variables has been chosen, the prediction for Z is determined via *linear regression*, i.e., as a linear combination $\sum_i \alpha_i X_i$ with appropriately fitted coefficients α_i. The past observed correlation between the X_i and Z is characterized in terms of the *covariance matrix* C between the X_i, and the vector $\mathbf{b}$ of covariances between Z and the X_i. The optimization problem can then be phrased as follows:

> Given C and $\mathbf{b}$, find a set S of size at most k so as to minimize the *mean squared prediction error* [10, 18] $\mathrm{Err}(Z,S) := \mathrm{E}\left[(Z - \sum_{i\in S}\alpha_i X_i)^2\right]$, where the α_i are the optimal regression coefficients specifically for the set S. Alternatively, maximize the *squared multiple correlation* [10] $R^2_{Z,S} := \frac{\mathrm{Var}[Z]-\mathrm{E}\left[(Z-\sum_{i\in S}\alpha_i X_i)^2\right]}{\mathrm{Var}[Z]}$.

The squared multiple correlation is proportional to the *error reduction* ($\mathrm{Var}[Z]$-$\mathrm{Err}(Z,S)$) of Z due to S. It can intuitively be viewed as the proportion of variance of the predictor variable that can be explained by the set of observation variables, and is often used in the statistics literature [4, 13, 26] to measure the quality of prediction during regression analysis. While other functions are also frequently used

to measure the accuracy of regression (such as conditional entropy [1], mean absolute error, etc.), we only focus on the above two objectives in this paper. Naturally, the optimization problems for both are equivalent at optimality, but the two functions differ significantly in terms of approximation and hardness results.

1.1 Our Results

The subset-selection problem is NP-hard in general. In fact, it is even NP-hard to decide whether any k-subset reduces the mean squared prediction error to 0 [8]. As a result, no multiplicative approximation guarantee for the $\mathrm{Err}(Z, S)$ objective is possible in general. Due to the highly non-linear nature of either objective function, a general approximation for $R^2_{Z,S}$ also seems very hard to obtain (though no approximation hardness result is known). It is therefore desirable to identify natural special cases amenable to efficient (approximation) algorithms for the $R^2_{Z,S}$ objective. To characterize such instances, we define the *covariance graph* on the node set $\{Z, X_1, \ldots, X_n\}$, with edges between any pair of variables with non-zero covariance. Our main contribution consists of algorithms for several classes of covariance graphs.

In order to prove these results, we first derive (in Section 3) a set of general perturbation results for the $R^2_{Z,S}$ objective. These results bound the effect of (small) perturbations in the covariance matrix C and covariance vector $\mathbf{b}$, so long as C is well-conditioned. One implication of these results is that correlations smaller than some ϵ can be ignored without changing the quality of the solution much. Thus, algorithms for a given class of covariance graphs yield algorithms with nearly the same guarantees for cases when the edges corresponding to covariance exceeding ϵ form a covariance graph in that class.

A second application of our perturbation results is an approximation guarantee, both for the $\mathrm{Err}(Z, S)$ and $R^2_{Z,S}$ objectives, for the widely used *Forward Regression* heuristic (also known as *Forward Selection*, and defined in Section 3.1), under the assumption that the X_i variables have small correlations. The guarantees for $\mathrm{Err}(Z, S)$ are similar to those in [14, 32] for slightly different greedy heuristics, under similar independence assumptions.

In terms of the covariance graph G induced by Z and the X_i, the previous result assumes G to be a star (except for edges of covariance at most ϵ) rooted at Z. We subsequently derive the following results for more general classes of covariance graphs. In all cases, our perturbation bounds yield approximate versions of these results when additional edges exist with covariances at most ϵ.

1. An FPTAS for the $R^2_{Z,S}$ objective, for the case when the subgraph of G induced only by the X_i has constant bandwidth, with Z being allowed arbitrary connections. This case has applications for time series analysis, when the X_i represent samples at regular time intervals, and correlations between the X_i exist only within sufficiently short time intervals (Section 4).

2. An exact algorithm for the case when G forms a tree. This is a significant and non-trivial generalization of the case of a star, which was studied previously (Section 5), and motivated by connections with tree models in machine learning.

3. An exact algorithm when G has a (known) independent set containing all but a constant number of variables (Section 6). This case is important for subset selection in practice when we have a set of independent variables, along with a constant number of aggregate statistics about these variables which can also be observed.

Finally, we depart from the graph-based view, and investigate two special types of covariance matrices:

1. If the variables can be embedded into a line such that the covariances decrease exponentially in the distance, then we obtain an exact algorithm (Section 7). This special case arises for multiplicative-decrease, random-increase processes over time. It also constitutes a first step toward a more thorough understanding of the optimization problem for variables embedded in some metric space with covariances monotonically decreasing in the distance, which has clear applications in modeling sensor networks.

2. If the instance has no "suppressor variables" (variables X_i which seem uncorrelated with Z until another variable X_j has been sampled), then we prove that Forward Regression gives a $(1 - 1/e)$ approximation for maximizing $R^2_{Z,S}$ (Section 8). This result is important in that it has been observed empirically in the past [5, 36] that much of the difficulty in subset selection results from such "suppressor variables". Our result confirms this empirical observation analytically.

1.2 Related Work

With the advent of sensor networks, much recent work has focused on tradeoffs between the accuracy in measurements and the energy expended in retrieving data. Deshpande et al. [9] use statistical models of sensor data to extrapolate sensor readings based on already collected sensor data, and reduce the number of sensors needed to answer a given query. Guestrin et al. [15] and Anstreicher et al. [1, 19] study related problems that deal with maximizing the entropy or mutual information of a subset: the aim is to find the most informative k-subset in a set of sensors, measured in terms of the joint entropy or mutual information of the subset of variables. Recently, Liaskovitis and Schurgers suggested a formulation essentially equivalent to the subset selection problem for choosing sensor sets to sample [20].

The formulation of the subset selection problem presented in this paper coincides with that in the statistics community [21]. Many heuristics have been proposed; the book by Miller [21] contains an extensive summary. Some of the well-known methods are Forward, Backward and Stepwise Regression, and Branch and Bound techniques. While there has not been any rigorous algorithmic analysis of Forward and Stepwise Regression, Couvreur and Bressler [7] analyzed Backward Regression and showed that if the optimal prediction error is smaller than a certain threshold, Backward Regression will select the optimal subset. However, [7] points out that calculating this threshold might itself be NP-hard.

Instead of restricting the number of variables sampled, other widely used regression models such as the Lasso method [31], and the Elastic Net [39], prescribe different constraints on the regression coefficient vectors. For instance, the Lasso method gives an upper bound on the l_1 norm of the regression coefficients vector (whereas the set size constraint can

be interpreted as a bound on the l_0 norm), which results in a convex optimization problem.

In the mathematics and signal processing communities, subset selection has been studied in the context of "sparse approximation". In this context, the problem consists of selecting a "sparse" subset from among a large dictionary ϕ of m vectors $\{\mathbf{d_i} \in \mathbb{R}^n \mid 1 \leq i \leq m\}$, whose linear combination best approximates a given signal vector $\mathbf{y} \in \mathbb{R}^n$ in the least square error sense. One formulation of this problem, equivalent to subset selection under approximation-preserving reductions, involves finding the best approximation of the input vector using at most k basis vectors from ϕ.

A paper by Gilbert et al. [14] was the first to rigorously prove approximation bounds of greedy solutions for the above-mentioned sparse approximation formulation, assuming nearly orthogonal dictionaries. They analyze a two-stage algorithm consisting of two slightly different greedy heuristics: "Matching Pursuit" and "Orthogonal Matching Pursuit". The approximation guarantees were subsequently improved by Tropp et al. [35, 32], the latter paper using just a single stage of greedy Orthogonal Matching Pursuit. Since the bounds of [14, 35, 32] are similar to our bounds for Forward Regression, we defer a precise statement and comparison to Section 3.1. Tropp [33] also presents a detailed study of variants of sparse approximation, and provides interesting results regarding the performance of greedy and convex relaxation methods when the dictionary is almost orthogonal.

Another commonly studied version of the sparse approximation problem in the signal processing community involves minimizing the number of basis vectors needed to achieve a desired estimation error ϵ in predicting the input vector. Assuming an $n \times m$ dictionary ϕ, the aim is to find a coefficient vector $\alpha_\mathbf{0} \in \mathbb{R}^m$ that minimizes $\|\alpha_\mathbf{0}\|_0$ subject to $\|y - \phi\alpha_\mathbf{0}\|_2 \leq \epsilon$. Natarajan [23] proved a weak approximation bound on the performance of a greedy algorithm for this problem. This problem has also been studied extensively in recent years in the context of sparse signal recovery from a set of noisy observations [3, 12, 11, 34, 37]. Almost all of these results use convex relaxation techniques to replace the l_0 norm constraint with an l_1 norm constraint, and solve the corresponding convex optimization problem. They then prove certain sparsity conditions under which the coefficient vector $\alpha_\mathbf{1}$ retrieved using l_1 relaxation can approximate the optimal coefficient vector $\alpha_\mathbf{0}$. In particular, recent results by Candes et al. [3] and Donoho [11] show that if the optimal coefficient vector is sufficiently sparse, then for all dictionary matrices satisfying an isometry condition, the $\alpha_\mathbf{1}$ vector obeys $\|\alpha_\mathbf{1} - \alpha_\mathbf{0}\|_2 \leq c\epsilon$, for a given constant c.

In the mathematical approximation theory community, Temlyakov [29, 30] analyzed convergence theorems to prove bounds on the power decay of approximation using sophisticated mathematical techniques in Hilbert and Banach spaces.

2. PRELIMINARIES

The goal is to estimate a *predictor variable* Z using a small subset of the *observation variables* $X_1, \ldots, X_n$. By appropriately scaling $X_i - \mathrm{E}[X_i]$, we can assume without loss of generality that all random variables have expectation 0 and variance 1.

$\mathrm{Var}[X_i]$, $\mathrm{Cov}[X_i, X_j]$ and $\rho(X_i, X_j)$ denote the variance, covariance and correlation of random variables, respectively. The matrix of covariances between X_i and X_j is denoted by C, so $c_{i,j} = \mathrm{Cov}[X_i, X_j]$. The vector $\mathbf{b}$ denotes the covariances between Z and the X_i, so $b_i = \mathrm{Cov}[Z, X_i]$. Recall [16] that a matrix C is a covariance matrix iff it is positive semi-definite. We use C_S to denote the submatrix with row and column set S, and $\mathbf{b_S}$ to denote the vector with only entries b_i for $i \in S$.

We denote the eigenvalues of an $n \times n$ symmetric matrix A as $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A)$. We use the following matrix and vector norms: $\|\mathbf{x}\|_2 = \sqrt{\sum_i (x_i)^2}$, $\|\mathbf{x}\|_\infty = \max_i |x_i|$, $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$ and $\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty$. It is easy to see that $\|A\|_2 = \lambda_1(A)$, and $\|A\|_\infty = \max_i \sum_j |a_{ij}|$. The *condition number* of the matrix A is defined as $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$.

We want to select and sample a set S of at most k variables X_i, and output a linear predictor $Z' = \sum_{i \in S} \alpha_i X_i$ of Z. The goal is to choose the set S and the coefficients α_i so as to minimize the *mean square prediction error* $\mathrm{E}\left[(Z - Z')^2\right]$, or, equivalently, maximize the *squared multiple correlation* $R^2_{Z,S} := \frac{\mathrm{Var}[Z] - \mathrm{E}\left[(Z-Z')^2\right]}{\mathrm{Var}[Z]}$ [10, 18]. The latter is a widely used measure for the goodness of a statistical fit; it captures the fraction of the variance of Z caused 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 notational convenience, we will frequently 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 α_i are well known to be $\mathbf{a_S} = (\alpha_i)_{i \in S} = C_S^{-1} \cdot \mathbf{b_S}$ (see, e.g., [18]), and hence $R^2_{Z,S} = \mathbf{b_S}^T (C_S^{-1})^T \mathbf{b_S}$. Thus, the subset selection problem can be phrased as follows:

Definition 2.1 (Subset Selection). *Given C, $\mathbf{b}$, and k, select a set S of k variables to minimize the mean square prediction error* $\mathrm{Err}(Z, S) = \mathrm{Var}[Z] - \mathbf{b_S}^T (C_S^{-1})^T \mathbf{b_S}$. *Equivalently, maximize* $R^2_{Z,S} = \mathbf{b_S}^T (C_S^{-1})^T \mathbf{b_S}$.[1]

For notational convenience, we define the following standard quantities (see [10]).

Definition 2.2 (Residual, Semipartial Correlation). *The random variable* $\mathrm{Res}(Z, S) = Z - \sum_{i \in S} \alpha_i X_i$ *is called the* residual *of Z with respect to the X_i for $i \in S$. It captures the part of Z not correlated with X_i for all $i \in S$.*

The semipartial correlation $\mathrm{Corr}(Z, X/Y) = \rho(Z, \mathrm{Res}(X, Y))$ *is the correlation between Z and the residual of X with respect to the variable Y.*

The following basic lemmas about residuals and squared multiple correlations will be particularly useful. Their proofs involve fairly standard manipulations, and are omitted from this version due to space constraints.

Lemma 2.3. *For any random variables Z, Y and a set of random variables S,*
$\mathrm{Res}(Z, S \cup \{Y\}) = \mathrm{Res}(\mathrm{Res}(Z, S), \{\mathrm{Res}(Y, S)\})$

Lemma 2.4. *For any random variables Z, Y and a set of random variables S,* $R^2_{Z,S \cup \{Y\}} = R^2_{Z,S} + R^2_{Z,\{\mathrm{Res}(Y,S)\}}$.

The subset selection problem from Definition 2.1 is equivalent to the sparse approximation problem [14, 33], via straightforward approximation preserving reductions. The reduction from sparse approximation to subset selection defines

[1] 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 generalized inverse.

the covariance $\mathrm{Cov}[X_i, X_j]$ to be equal to the inner product of the ith and j^{th} vectors of the dictionary. This technique is also useful in practice for the best estimate of the covariance matrix from empirical joint observations of the observation variables X_i. The other direction defines the dictionary through the Cholesky Decomposition of the covariance matrix C. Details are deferred to the full version of this paper due to space constraints.

As shown in [8, 23, 22] (where the results are phrased equivalently in terms of sparse approximation), the subset selection problem is NP-complete, and the minimization version can in general not be approximated to within any constant. In fact, Muthukrishnan [22] shows that unless $\mathrm{P} \subseteq \mathrm{DTIME}[n^{O(\log\log n)}]$, there is no polynomial-time approximation algorithm approximating the error within any constant α, while using at most $O(k \log n)$ random variables.

We therefore focus on natural special classes of the subset selection problem. Specifically, we propose restricting the graph structure of the covariance matrix to identify more tractable scenarios. Formally, we define:

Definition 2.5 (Covariance Graphs). *For any $\epsilon \geq 0$, the* covariance graph *$G_\epsilon(C, \mathbf{b})$ of $C, \mathbf{b}$ is the graph with node set $\{X_0 := Z, X_1, \ldots, X_n\}$ and edges between any pair of variables X_i, X_j with $\mathrm{Cov}[X_i, X_j] \geq \epsilon$. For simplicity, we write $G(C, \mathbf{b}) := G_0(C, \mathbf{b})$.*

The covariance graph on observation variables only is denoted by $\tilde{G}_\epsilon(C)$, with node set $\{X_1, \ldots, X_n\}$, and otherwise defined analogously to $G_\epsilon(C, \mathbf{b})$.

3. MATRIX PERTURBATION BOUNDS

In this section, we employ matrix perturbation bounds to show that for sufficiently well-conditioned covariance matrices C, small perturbations to the matrix only cause small changes in the $R^2_{Z,S}$ objective function.

Lemma 3.1. *Assume that the $k \times k$ covariance matrix C is non-singular, and let κ be its condition number. Let E be a $k \times k$* error matrix *whose largest entry in absolute value is at most $\delta \leq \frac{1}{4\kappa k}$. Then, $\frac{|\mathbf{b}^T C^{-1}\mathbf{b} - \mathbf{b}^T (C+E)^{-1}\mathbf{b}|}{|\mathbf{b}^T C^{-1}\mathbf{b}|} \leq \frac{4}{3}\kappa^2 k\delta$.*

Thus, so long as the covariances between any pair of observation variables are changed by no more than a small δ, the relative change in the error reduction for a set S of size k is at most $\frac{4}{3}\kappa^2 k\delta$. There are simple examples that show that small perturbations can cause large changes in the error reduction if the matrix is nearly singular; hence, the dependence of the bound on κ is necessary.

Proof. Using linearity, the Cauchy-Schwartz Inequality, and the definition of an induced matrix norm, we first obtain the bound $\frac{|\mathbf{b}^T C^{-1}\mathbf{b} - \mathbf{b}^T (C+E)^{-1}\mathbf{b}|}{|\mathbf{b}^T C^{-1}\mathbf{b}|} \leq \frac{\|\mathbf{b}\|_2^2 \|(C+E)^{-1} - C^{-1}\|_2}{|\mathbf{b}^T C^{-1}\mathbf{b}|}$.

By considering the representation of $\mathbf{b}$ in terms of an orthonormal set of eigenvectors of C^{-1}, it is not difficult to prove that $|\mathbf{b}^T C^{-1}\mathbf{b}| \geq \frac{1}{\lambda_1(C)}\|\mathbf{b}\|_2^2$. Hence, we have that $\frac{|\mathbf{b}^T C^{-1}\mathbf{b} - \mathbf{b}^T (C+E)^{-1}\mathbf{b}|}{|\mathbf{b}^T C^{-1}\mathbf{b}|} \leq \lambda_1 \|(C+E)^{-1} - C^{-1}\|_2$.

To bound the right-hand side, we can invoke a well-known theorem from perturbation theory of matrix inverses, Theorem III.2.5 from [28]. Assuming that $\|C^{-1}E\|_2 < 1$, the theorem, submultiplicativity of matrix norms, and the fact that $\|C^{-1}\|_2 = 1/\lambda_k$, together imply that

$$\begin{aligned} \lambda_1 \|(C+E)^{-1} - C^{-1}\|_2 &\leq \lambda_1 \tfrac{\|C^{-1}E\|_2}{1 - \|C^{-1}E\|_2} \cdot \|C^{-1}\|_2 \\ \leq \lambda_1 \tfrac{\|C^{-1}\|_2 \|E\|_2}{1 - \|C^{-1}\|_2 \|E\|_2} \cdot \|C^{-1}\|_2 &\leq \kappa^2 \tfrac{\|E\|_2}{\|C\|_2 - \kappa\|E\|_2}. \end{aligned}$$

If each entry of E is at most δ, then $\|E\|_2 \leq \delta k$. Also, $\|C\|_2 \geq 1$, because the length of $C \cdot (1, 0, 0, \ldots, 0)$ is at least 1. Substituting these now proves the lemma. ■

Applying the above lemma to the case where C is the identity matrix, i.e., none of the observation variables are linearly correlated, gives us the following corollary:

Corollary 3.2. *If k normalized random variables have pairwise covariances of at most $\delta \leq \frac{1}{4k}$, then the error reduction in Z from this set satisfies $\frac{|R^2_{Z,X_1,\ldots,X_k} - \sum_{i=1}^k b_i^2|}{\sum_{i=1}^k b_i^2} \leq \frac{4}{3}\delta k$.*

Using a similar analysis, we also obtain a perturbation bound for the covariance vector $\mathbf{b}$.

Lemma 3.3. *Assume that the $k \times k$ covariance matrix C is non-singular, and let $\kappa = \kappa(C) > 0$ be its condition number. Let $\mathbf{e}$ be a k-dimensional* error vector, *such that $\frac{\|\mathbf{e}\|_2}{\|\mathbf{b}\|_2} \leq \delta$, where $\delta \leq \frac{1}{4\kappa k}$. Then, $\frac{|(\mathbf{b}+\mathbf{e})^T C^{-1}(\mathbf{b}+\mathbf{e}) - \mathbf{b}^T C^{-1}\mathbf{b}|}{|\mathbf{b}^T C^{-1}\mathbf{b}|} \leq 3\kappa\delta$.*

Proof. Using linearity, the Cauchy-Schwartz Inequality, and the definition of an induced matrix norm, we first obtain

$$\frac{|(\mathbf{b}+\mathbf{e})^T C^{-1}(\mathbf{b}+\mathbf{e}) - \mathbf{b}^T C^{-1}\mathbf{b}|}{|\mathbf{b}^T C^{-1}\mathbf{b}|} \leq \frac{\|2\mathbf{b}+\mathbf{e}\|_2 \|C^{-1}\|_2 \|\mathbf{e}\|_2}{|\mathbf{b}^T C^{-1}\mathbf{b}|}.$$

By considering the representation of $\mathbf{b}$ in terms of an orthonormal set of eigenvectors of C^{-1}, it can be seen that $|\mathbf{b}^T C^{-1}\mathbf{b}| \geq \frac{1}{\lambda_1}\|\mathbf{b}\|_2^2$. Hence, $\frac{|(\mathbf{b}+\mathbf{e})^T C^{-1}(\mathbf{b}+\mathbf{e}) - \mathbf{b}^T C^{-1}\mathbf{b}|}{|\mathbf{b}^T C^{-1}\mathbf{b}|} \leq \frac{\kappa\|2\mathbf{b}+\mathbf{e}\|_2 \|\mathbf{e}\|_2}{\|\mathbf{b}\|_2^2} \leq \frac{3\kappa\|\mathbf{e}\|_2}{\|\mathbf{b}\|_2} \leq 3\kappa\delta$. ■

3.1 Forward Regression

We now use the perturbation results above to give an analysis for the commonly used *Forward Regression* algorithm (also called *Forward Selection*). Forward Regression is frequently used in the social sciences; hence, proving guarantees for its performance is important. If the covariance terms are sufficiently small, then Forward Regression yields good approximation guarantees for both $\mathrm{Err}(Z, S)$ and $R^2_{Z,S}$.

Definition 3.4 (Forward Regression Heurestic). *Select a set S of size k iteratively as follows: (1) Initialize $S_0 = \emptyset$. (2) In each step $i+1$, select a variable X_j minimizing $\mathrm{Err}(Z, S_i \cup \{X_j\})$. (3) Output S_k.*

We first prove a bound on the approximation of $R^2_{Z,S}$, and then derive a theorem about the approximation of $\mathrm{Err}(Z, S)$.

Theorem 3.5. *If $\mathrm{Cov}[X_i, X_j] \leq \delta < \frac{1}{4k}$ for all i, j, the Forward Regression heurestic reduces the error by at least $R^2_{Z,S} \geq (1 - 4\delta k) \cdot R^2_{Z,S^*}$, where S^* is the optimum solution.*

Proof. We assume without loss of generality that $b_1 \geq b_2 \geq \ldots \geq b_n$. Let $S = \{Y_1, \ldots, Y_k\}$ be the variables chosen by the Forward Regression heuristic. We will show that $R^2_{Z,S} \geq (1 - \frac{8}{3}\delta k)\sum_{i=1}^k b_i^2$. Using Corollary 3.2, applied to S^*, and combining it with the optimality of $\sum_{i=1}^k b_i^2$ for the identity matrix, this implies that $R^2_{Z,S} \geq (1 - \frac{8}{3}\delta k)(1 - \frac{4}{3}\delta k)R^2_{Z,S^*} \geq (1 - 4\delta k)R^2_{Z,S^*}$, and hence the theorem.

We prove the inequality by induction over the k iterations of the greedy algorithm. Clearly, the first variable chosen by Forward Regression is X_1, so the base case holds. Assume that $R^2_{Z,S_j} \geq (1-\frac{8}{3}\delta j)\sum_{i=1}^{j} b_i^2$ after iteration j. Because the variance of Z is normalized to 1, and by optimality of the choice made by Forward Regression, we know that $R^2_{Z,S_{j+1}} = R^2_{Z,S_j} + R^2_{Z,\mathrm{Res}(Y_{j+1},S_j)} \geq R^2_{Z,S_j} + R^2_{Z,\mathrm{Res}(X_{j+1},S_j)}$.

Then, by using the induction hypothesis, and the fact that $\mathrm{Var}[\mathrm{Res}(X_{j+1},S_j)] \leq 1$, we bound $R^2_{Z,S_j} + R^2_{Z,\mathrm{Res}(X_{j+1},S_j)} \geq (1-\frac{8}{3}\delta j)\sum_{i=1}^{j} b_i^2 + R^2_{Z,\mathrm{Res}(X_{j+1},S_j)} \geq (1-\frac{8}{3}\delta j)\sum_{i=1}^{j} b_i^2 + \mathrm{Cov}[Z,\mathrm{Res}(X_{j+1},S_j)]^2$.

Let $\mathbf{b}' = [c_{j+1,i}]_{i\in S_j}$ be the vector of covariances between X_{j+1} and the variables X_i for $i \in S_j$, and C_{S_j} the submatrix of the covariance matrix induced by the elements of S_j. Let $\mathbf{a} = C_{S_j}^{-1}\mathbf{b}'$ be the vector of optimal regression coefficients in $X_{j+1} = \sum_{i=1}^{j} a_i Y_i + \mathrm{Res}(X_{j+1},S_j)$. In order to show that the a_i must be small, we use another matrix perturbation theorem, Theorem III.2.11 from [28]. Let $E = C_{S_j} - I$. Because each entry of E is at most $\delta < \frac{1}{4j}$, the theorem, applied to the identity matrix, implies that $\frac{\|\mathbf{a}-\mathbf{b}'\|_\infty}{\|\mathbf{b}'\|_\infty} = \frac{\|(I+E)^{-1}\mathbf{b}' - I^{-1}\mathbf{b}'\|_\infty}{\|I^{-1}\mathbf{b}'\|_\infty} \leq \frac{\|E\|_\infty}{1-\|E\|_\infty} \leq \frac{j\delta}{1-j\delta} \leq \frac{1}{3}$.

By the triangle inequality, $\|\mathbf{a}\|_\infty \leq \frac{4}{3}\|\mathbf{b}'\|_\infty \leq \frac{4}{3}\delta$. This implies $|a_i| \leq \frac{4}{3}\delta$, for all i. Expanding $\mathrm{Cov}[Z,\mathrm{Res}(X_{j+1},S_j)]^2$, using the fact that $\sum_{i=1}^{j}\mathrm{Cov}[Z,Y_i] \leq \sum_{i=1}^{j}\mathrm{Cov}[Z,X_i]$, and substituting the bound on $|a_i|$, we obtain that

$$\begin{aligned}
& \mathrm{Cov}[Z,\mathrm{Res}(X_{j+1},S_j)]^2 \\
= \;& (\mathrm{Cov}[Z,X_{j+1}] - \textstyle\sum_{i=1}^{j} a_i \mathrm{Cov}[Z,Y_i])^2 \\
\geq \;& \mathrm{Cov}[Z,X_{j+1}]^2 - 2\mathrm{Cov}[Z,X_{j+1}]\textstyle\sum_{i=1}^{j}|a_i||\mathrm{Cov}[Z,Y_i]| \\
\geq \;& b_{j+1}^2 - \tfrac{8}{3}\delta\cdot b_{j+1}\textstyle\sum_{i=1}^{j}|b_i|.
\end{aligned}$$

Using the sorting of the X_i, we have that $b_{j+1} \leq b_i$ for all $i \leq j$, so the above can be further bounded as $b_{j+1}^2 - \frac{8}{3}\delta\sum_{i=1}^{j} b_i^2$. Finally, we can substitute this bound back into our lower bound for $R^2_{Z,S_{j+1}}$, obtaining that

$$\begin{aligned}
R^2_{Z,S_{j+1}} \;&\geq\; (1-\tfrac{8}{3}\delta j)\textstyle\sum_{i=1}^{j} b_i^2 + b_{j+1}^2 - \tfrac{8}{3}\delta\sum_{i=1}^{j} b_i^2 \\
&\geq\; (1-\tfrac{8}{3}\delta(j+1))\textstyle\sum_{i=1}^{j+1} b_i^2.
\end{aligned}$$

This completes the inductive proof, and hence the proof of the theorem. ■

Theorem 3.5 is one of the key ingredients in the proof of the following theorem about the approximation guarantee of Forward Regression with respect to the mean squared error $\mathrm{Err}(Z,S)$.

THEOREM 3.6. *Assume that* $\mathrm{Cov}[X_i,X_j] \leq \delta < \frac{1}{6k}$ *for all* i,j. *Let* S *be the set selected by Forward Regression, and* S^* *the optimum solution. Then,* $\mathrm{Err}(Z,S) \leq (1+16(k+1)^2\delta)\cdot\mathrm{Err}(Z,S^*)$.

We begin with two useful lemmas. The first (whose proof is similar to that of Theorem 3.5, and deferred to the full version due to space constraints) shows that if there is a variable with large covariance with Z that is not included in the optimum solution, then the variance of Z itself cannot be too large (or the error of the optimum solution is still quite large).

LEMMA 3.7. *Suppose that* $X_1 \notin S^*$. *If* $\delta \leq \frac{1}{4k}$, *then* $\mathrm{Var}[Z] \leq (4k+1)\cdot\mathrm{Err}(Z,S^*)$.

We can use this lemma to derive a corollary showing that at the time the Forward Regression heuristic picks a "wrong" variable for the first time, its solution cannot be too far from the overall optimum.

COROLLARY 3.8. *Let* S_i *be the set chosen by Forward Regression after iteration* i, *and* t *be the largest iteration number such that* $S_t \subseteq S^*$. *If* $\delta \leq \frac{1}{6k}$, *then* $\mathrm{Err}(Z,S_t) \leq (4k+1)\mathrm{Err}(Z,S^*)$.

Proof. Consider replacing Z with $Z' = \frac{\mathrm{Res}(Z,S_t)}{\sqrt{\mathrm{Var}[\mathrm{Res}(Z,S_t)]}}$, and each $X_i \notin S_t$ with $X_i' = \frac{\mathrm{Res}(X_i,S_t)}{\sqrt{\mathrm{Var}[\mathrm{Res}(X_i,S_t)]}}$. That is, the component correlated with S_t is removed from all variables, and the variables are then renormalized to have variance 1. Now, consider finding the optimal subset of size $k-t$ for this new problem. We will show that $\mathrm{Cov}[X_i',X_j'] \leq \frac{3}{2}\delta \leq \frac{1}{4k}$. By definition of Forward Regression, the variable picked in iteration $t+1$ maximizes $\mathrm{Cov}[Z',X_i']$, and by choice of t, that variable is not in S^*. But $S^* \setminus S_t$ must be an optimal solution for the new problem, so we can apply Lemma 3.7 directly to the new problem, and obtain the desired bound.

For any variable $X_i \notin S_t$, Corollary 3.2 and the fact that $\mathrm{Var}[X_i] = 1$ imply that $\mathrm{Var}[\mathrm{Res}(X_i,S_t)] = 1 - R^2_{X_i,S_t} \geq 1-(1+\frac{4}{3}t\delta)\cdot\delta^2 t \geq 1-\frac{1}{3}\delta$.

For any pair of variables $X_i, X_j \notin S_t$, using the definition of residuals, it is easy to prove $\mathrm{Cov}[\mathrm{Res}(X_i,S_t),\mathrm{Res}(X_j,S_t)] = \mathrm{Cov}[X_i,X_j] - \mathrm{Cov}[X_i,X_j - \mathrm{Res}(X_j,S_t)] = \mathrm{Cov}[X_i,X_j] - \mathrm{Cov}[X_i - \mathrm{Res}(X_i,S_t), X_j - \mathrm{Res}(X_j,S_t)]$.

Using the Cauchy Schwarz Inequality and Corollary 3.2, we can bound the second term as $|\mathrm{Cov}[X_i - \mathrm{Res}(X_i,S_t), X_j - \mathrm{Res}(X_j,S_t)]| \leq \sqrt{R^2_{X_i,S_t}R^2_{X_j,S_t}} \leq (1+\frac{4}{3}t\delta)\delta^2 t \leq \frac{1}{3}\delta$.

Thus, $\mathrm{Cov}[\mathrm{Res}(X_i,S_t),\mathrm{Res}(X_j,S_t)] \leq \frac{4}{3}\delta$, and after normalizing the variances to 1 again, the new covariances are at most $\mathrm{Cov}[X_i',X_j'] \leq (\frac{4}{3}\delta)/(1-\frac{1}{3}\delta) \leq \frac{3}{2}\delta$. ■

Using Theorem 3.5 and Corollary 3.8, we can now complete the proof of Theorem 3.6.

Proof of Theorem 3.6. We analyze Forward Regression in two stages. Let t be the latest iteration such that $S_t \subseteq S^*$. Define Z' and X_i' as in the proof of Corollary 3.8. Consider the remaining $k-t$ iterations of Forward Regression with Z' and X_i'. Because $S^* \setminus S_t$ is the optimal solution for this new problem, Theorem 3.5, applied to the solution $S \setminus S_t$ found by Forward Regression, implies (using $\mathrm{Var}[Z'] = 1$) that

$$\begin{aligned}
\mathrm{Var}[\mathrm{Res}(Z',S\setminus S_t)] \;&\leq\; \mathrm{Var}[Z'] - (1-4\delta k)R^2_{Z',S^*\setminus S_t} \\
&\leq\; \mathrm{Var}[\mathrm{Res}(Z',S^*\setminus S_t)] + 4\delta k.
\end{aligned}$$

The identities $\mathrm{Res}(Z,S) = \mathrm{Res}(Z',S\setminus S_t)\cdot\sqrt{\mathrm{Var}[\mathrm{Res}(Z,S_t)]}$ and $\mathrm{Res}(Z,S^*) = \mathrm{Res}(Z',S^*\setminus S_t)\cdot\sqrt{\mathrm{Var}[\mathrm{Res}(Z,S_t)]}$ both follow from the definition of Z', and using Corollary 3.8 imply that

$$\begin{aligned}
\mathrm{Var}[\mathrm{Res}(Z,S)] \;&\leq\; \mathrm{Var}[\mathrm{Res}(Z,S^*)] + 4\delta k\mathrm{Var}[\mathrm{Res}(Z,S_t)] \\
&\leq\; \mathrm{Var}[\mathrm{Res}(Z,S^*)] + 4\delta k(4k+1)\mathrm{Err}(Z,S^*) \\
&\leq\; (1+16(k+1)^2\delta)\mathrm{Err}(Z,S^*),
\end{aligned}$$

completing the proof of the theorem. ■

Theorem 3.6 proves similar guarantees about Forward Regression as [14, 35, 32] proved about different greedy heuristics. Using our notation, the result of [14] for a two-stage

greedy algorithm of Matching Pursuit and Orthogonal Matching Pursuit states that if the maximum covariance between pairs of observation variables is $\delta < \frac{1}{32k}$, the two-stage algorithm finds a k-subset S with mean-square prediction error $\mathrm{Err}(Z,S) \leq (1+2064k^2\delta)\cdot \mathrm{Err}(Z,S^*)$. Tropp et al. [35] improved this bound to $\mathrm{Err}(Z,S) \leq (1+\frac{2k^2\delta}{(1-2k\delta)^2})\cdot \mathrm{Err}(Z,S^*)$, whenever $\delta < \frac{1}{2k}$. Tropp [32] subsequently showed a similar guarantee for the Orthogonal Matching Pursuit algorithm without the other stage. His results prove that the mean-square prediction error is $\mathrm{Err}(Z,S) \leq (1+6k)\cdot \mathrm{Err}(Z,S^*)$, whenever $\delta < \frac{1}{3k}$.

4. LOW BANDWIDTH GRAPHS

In this section, we study the special case where $\tilde{G}(C)$ has constant bandwidth β. That is, the variables can be ordered such that $\mathrm{Cov}[X_i, X_j] = 0$ whenever $|j-i| > \beta$. Under the assumption that C has polynomially bounded condition number, we give an FPTAS for maximizing the error reduction, $R^2_{Z,S}$. Combining it with our matrix perturbation bound from Lemma 3.1 also yields approximation algorithms for the case where $\tilde{G}(C)$ is sufficiently close to having constant bandwidth. The low-bandwidth case is important for its application in two practical cases. It can model random variables of a time series where the temporal correlations between variables are significant only within a small time interval. Additionally, it can model spatial correlations between sensor variables placed on a line.

Using an algorithm of Saxe [27], an ordering with bandwidth β can be found in polynomial time $O(n^\beta)$ if it exists. We assume from now on that the variables are ordered accordingly.

THEOREM 4.1. *If $\tilde{G}(C)$ has bounded bandwidth β, then there is an FPTAS for $R^2_{Z,S}$. For any $\epsilon > 0$, we can obtain a solution S in time $O(n(k/\epsilon)^{\beta^2}\beta^{2\beta^2}\kappa^{2\beta^2}(1+2\log(1/\rho_{\min}))^\beta)$, guaranteeing that $R^2_{Z,S} \geq (1-\epsilon)\cdot R^2_{Z,S^*}$, where S^* is the optimum solution. Here, $\rho_{\min}$ is the smallest non-zero correlation value of any X_i with Z.*

Proof. Given a desired approximation guarantee $(1-\epsilon)$, we first define $\delta := \frac{\epsilon}{6ek(\beta-1)^2\kappa^2}$. Given a (for now) arbitrary matrix $P = (p_{i,j}) \in \mathbb{R}^{(\beta-1)\times(\beta-1)}$ and vector $\mathbf{q} \in \mathbb{R}^{\beta-\mathbf{1}}$, we define $f(a,s,P,\mathbf{q})$ to be the maximum R^2 value for predicting Z using a size-s subset of variables $Y_1,\ldots,Y_{n-a+1}$, which are defined in terms of their covariance matrix C' (with entries $c'_{i,j}$) and covariance vector $\mathbf{b}'$ (with entries b'_i) with Z as follows: For $1 \leq i,j \leq \beta-1$, the covariances between Y_i and Y_j are given by $p_{i,j}$, and the covariances between Y_i and Z are given by q_i. For all other i,j, we have $c'_{i,j} = c_{a+i-1,a+j-1}$, and $b'_i = b_{a+i-1}$. Then, if r is the smallest index of a variable $Y_1,\ldots,Y_{n-a+1}$ not selected in the optimal s-subset, we have the recurrence:

$$f(a,s,P,\mathbf{q}) = \begin{cases} 0, \text{ if } a > n \text{ or if } s = 0; \\ R^2(Z, Y_1,\ldots,Y_{n-a+1}), \text{ if } n-a+1 \leq s; \\ \max_{r \leq n-a+1} R^2(Z, Y_1, \ldots Y_{r-1}) \\ +f(r+a, s-(r-1), P^{(r)}, \mathbf{q^{(r)}}) \text{ otherwise;} \end{cases}$$

where the matrix $P^{(r)}$ has as its $(i,j)^{th}$ entry a value equal to $\mathrm{Cov}[\mathrm{Res}(Y_{r+i}, \{Y_1,\ldots,Y_{r-1}\}), Y_{r+j}]$, and the vector $\mathbf{q^{(r)}}$ has as its ith component a value equal to $\mathrm{Cov}[Z, \mathrm{Res}(Y_{r+i},)\{Y_1,\ldots,Y_{r-1}\}]$. The last equation in the recurrence follows from Lemma 2.4 and the fact that the problem instance has bandwidth β.

The optimal error reduction in Z using k variables selected from $X_1, X_2, \ldots, X_n$ is then $f(1,k,P_0,\mathbf{q_0})$, where P_0 is the leading $(\beta-1)\times(\beta-1)$ submatrix of C, and $\mathbf{q_0}$ the vector of covariances of the first $\beta-1$ variables X_i with Z.

Since we cannot store the optimum f values for all parameter settings (from a continuous space), we discretize the allowable entries of P and $\mathbf{q}$ to create approximations P' and $\mathbf{q}'$, and use perturbation bounds to analyze the amount of error introduced. Each entry of the $(\beta-1)\times(\beta-1)$ leading submatrix of C' is rounded to the nearest multiple of δ. Thus, for each entry, there are at most $\frac{1}{\delta}$ possible values. By Lemma 4.3 (proved below), the condition number of C' in each iteration is at most that of C, and an extension of Lemma 3.1 thus implies a relative error of at most $\frac{4}{3}(\beta-1)\kappa^2\delta$ due to rounding the entries of P in each iteration.

Similarly, we round each of the first $\beta-1$ entries of $\mathbf{b}'$ as follows: Without loss of generality, the Y_i variables have been normalized to have a variance of 1. Let $\rho_{\min}$ be the smallest non-zero correlation value of any of the variables Y_i with Z. For each of the entries q_i, we distinguish whether they fall into the interval $[0,\rho_{\min}]$, in which case we round q_i to the closest multiple of $\delta\rho_{\min}$, or into the range $[\rho_{\min},1]$, in which case we round to the nearest value $(1+\delta)^i\rho_{\min}$ for $i = 0,\ldots,\log_{1+\delta}\frac{1}{\rho_{\min}}$. So the total number of possible rounded values for each entry is $\frac{1}{\delta} + \log_{1+\delta}\frac{1}{\rho_{\min}} \leq \frac{1}{\delta}\cdot(1+2\log(1/\rho_{\min}))$.

If we round just one entry q_i, then the resulting error vector $\mathbf{e}$ due to rounding in both cases satisfies $\frac{\|\mathbf{e}\|_2}{\|\mathbf{q}\|_2} \leq \delta$, i.e., it has small relative norm. Then, using Lemma 3.3, the relative perturbation error in f due to rounding a single q_i parameter is at most $3\kappa\delta$. Thus, the relative error due to discretizing all the entries of $\mathbf{q}$ in each iteration is at most $3(\beta-1)\kappa\delta$. Combining the relative errors introduced due to all of the rounding, we obtain the bound $|\frac{f(a,s,P',\mathbf{q}')-f(a,s,P,\mathbf{q})}{f(a,s,P,\mathbf{q})}| \leq 3(\beta-1)^2\kappa^2\delta$. Since there are at most k iterations, the total relative error in the worst case is $(1+3(\beta-1)^2\kappa^2\delta)^k - 1 \leq 3ek(\beta-1)^2\kappa^2\delta = \epsilon/2$, because $3(\beta-1)^2\kappa^2\delta \leq 1/k$. This implies an approximation guarantee of $1-\epsilon$, as desired.

Since $1/\delta = O((k\beta^2\kappa^2)/\epsilon)$, the table for the dynamic program has size $O(nk\cdot\frac{1}{\delta^{(\beta-1)^2}}\cdot(\frac{1}{\delta}\cdot(1+2\log(1/\rho_{\min})))^{\beta-1})$ $= O(n(k/\epsilon)^{\beta^2}\beta^{2\beta^2}\kappa^{2\beta^2}(1+2\log(1/\rho_{\min}))^\beta)$, and the total time to fill it is thus also polynomial in n,k, and $1/\epsilon$ for any fixed β. Notice that $\log(1/\rho_{\min})$ depends on the input only polynomially, not pseudo-polynomially. Thus, we obtain an FPTAS whenever $\kappa = O(\mathrm{poly}(n))$. ■

Combining the theorem with our matrix perturbation bound from Lemma 3.1, we obtain an approximation algorithm for the subset selection problem when the covariance matrix is sufficiently close to having low bandwidth and is well conditioned.

COROLLARY 4.2. *If the covariance graph $\tilde{G}_\delta(C)$ has bandwidth β, and $\delta \leq \frac{1}{4\kappa k}$, then for any $\epsilon > 0$, there is a polynomial-time approximation algorithm for $R^2_{Z,S}$, guaranteeing that $R^2_{Z,S} \geq (1-\frac{8}{3}\kappa(C)^2k\delta)(1-\epsilon)\cdot R^2_{Z,S^*}$, where S^* is the optimum solution.*

LEMMA 4.3. *Let $\kappa(C)$ be the condition number of the covariance matrix C of n random variables $X_1, X_2, \ldots, X_n$, and $\kappa(C')$ be the condition number for the $(n-1)\times(n-1)$ covariance matrix C' of the $n-1$ random variables* $\mathrm{Res}(X_2, X_1)$, $\mathrm{Res}(X_3, X_1), \ldots, \mathrm{Res}(X_n, X_1)$. *Then $\kappa(C) \geq \kappa(C')$.*

Proof. Let λ_i resp. λ'_i denote the eigenvalues of C resp. C'. By definition of the induced matrix norm for a symmetric matrix, $\kappa(C) = \frac{\lambda_1}{\lambda_n}$, and $\kappa(C') = \frac{\lambda'_1}{\lambda'_{n-1}}$.

We first prove that $\lambda_1 \geq \lambda'_1$. Consider the $(n-1) \times (n-1)$ covariance matrix C'' for $X_2, X_3, \ldots, X_n$. Since C'' is formed by removing the first row and column of C, $\lambda_1 \geq \lambda''_1$. Now $c'_{i,i} = \mathrm{Var}[\mathrm{Res}(X_{i+1}, X_1)] = c_{i+1,i+1} - \frac{c^2_{i+1,1}}{c_{1,1}}$, and $c'_{i,j} = \mathrm{Cov}[\mathrm{Res}(X_{i+1}, X_1), \mathrm{Res}(X_{j+1}, X_1)] = c_{i+1,j+1} - \frac{c_{i+1,1}c_{j+1,1}}{c_{1,1}}$. Hence we obtain that $C'' = C' + D$, where we define $D = \frac{1}{c_{1,1}}[c''_{2,1}, c''_{3,1}, \ldots, c''_{n,1}] \cdot [c''_{2,1}, c''_{3,1}, \ldots, c''_{n,1}]^T$.

Clearly, D is positive semidefinite. Also, C' is a valid covariance matrix and hence is positive semidefinite. Therefore, by using Weyl's Theorem [16], we get $\lambda''_1 \geq \lambda'_1$, and hence $\lambda_1 \geq \lambda'_1$.

To prove $\lambda_n \leq \lambda'_{n-1}$, let $\mathbf{e}' = [e'_1, \ldots, e'_{n-1}]^T$ be the eigenvector of C' corresponding to λ'_{n-1}, and consider the n-dimensional vector $\mathbf{e} = [-\frac{1}{c_{1,1}} \sum_{i=2}^{n} e'_{i-1} c_{1,i}, e'_1, e'_2, \ldots, e'_{n-1}]^T$. Then, we have that $C \cdot \mathbf{e} = [0, \lambda'_{n-1}e'_1, \lambda'_{n-1}e'_2, \ldots, \lambda'_{n-1}e'_{n-1}]^T = \lambda'_{n-1}[0, e'_1, e'_2, \ldots e'_{n-1}]^T$. Thus, $\|C \cdot \mathbf{e}\|_2 \leq \lambda'_{n-1}\|\mathbf{e}\|_2$, which implies that $\lambda_n \leq \lambda'_n$. Hence $\kappa(C) \geq \kappa(C')$. ■

5. TREE COVARIANCE GRAPHS

In this section, we consider the case that $G(C, \mathbf{b})$ is a tree, and prove the following theorem.

THEOREM 5.1. *If the covariance graph $G(C, \mathbf{b})$ is a tree with maximum degree d, then the optimum k-subset for regression can be found in time $O(k^2 nd)$.*

Analogous to Corollary 4.2, this theorem can be combined with Lemma 3.1 to give an approximation algorithm if $G_\delta(C, \mathbf{b})$ forms a tree, i.e., the covariance matrix defines "nearly a tree" (assuming C is well-conditioned). Notice that a tree covariance matrix is a significant extension of the case when $G_\epsilon(C, \mathbf{b})$ forms a star, the only case for which provable guarantees were known in the past.

Proof. For ease of notation, we write $X_0 := Z$. We consider the tree as rooted at X_0. Let $S_{i,k}$ be the k-subset minimizing $\mathrm{Err}(X_i, S)$ among subsets using only variables in the subtree rooted at X_i, and $P(X_i, k) = \mathrm{Err}(X_i, S_{i,k})$ the error it achieves. For convenience, set $P(X_i, -1) = \infty$.

Let the children of X_i be $X^i_1, X^i_2, \ldots X^i_p$, for $p \leq d$. W.l.o.g., the optimum solution includes $X^i_1, \ldots X^i_m$. Then, it cannot include any nodes X_j in the subtrees rooted at $X^i_{m+1}, \ldots, X^i_p$. For an easy calculation of matrix inverses shows that if the node X_j is not connected to X_i with a path in the subgraph induced by S, then $\mathrm{Err}(X_i, S) = \mathrm{Err}(X_i, S \cup \{X_j\})$. Let R_j for $j = 1, \ldots, m$ denote the sets of nodes in the subtree rooted at X^i_j that are part of $S_{i,k}$, except X^i_j itself. We will show that if $X^i_j = X_{i'}$, and R_j has size k_j, then $R_j = S_{i',k_j}$. In words, the variables below $X_{i'}$ that are best for predicting $X_{i'}$ are also best for predicting X_i.

By applying Lemma 2.3, and then using the fact that Z is not linearly correlated with any $X_{i'} \in R_j$, and any $X_{i'}$ in R_j is not correlated with any variables except possibly X^i_j and others in R_j, we obtain that

$$\begin{aligned}
\mathrm{Res}(X_i, S_{i,k}) &= \mathrm{Res}(X_i, \textstyle\bigcup_{j=1}^m (R_j \cup \{X^i_j\})) \\
&= \mathrm{Res}(\mathrm{Res}(X_i, \textstyle\bigcup_j R_j), \{\mathrm{Res}(X^i_1, \textstyle\bigcup_j R_j), \\
&\quad \ldots, \mathrm{Res}(X^i_m, \textstyle\bigcup_j R_j)\}) \\
&= \mathrm{Res}(X_i, \{\mathrm{Res}(X^i_1, R_1), \ldots, \mathrm{Res}(X^i_m, R_m)\}).
\end{aligned}$$

Using the fact that $\mathrm{Cov}[\mathrm{Res}(X^i_j, R_j), \mathrm{Res}(X^i_{j'}, R_{j'})] = 0$ for $j \neq j'$, and applying Lemma 2.4, we have that $P(X_i, k) = \mathrm{Var}[\mathrm{Res}(X_i, S_{i,k})] = \mathrm{Var}[X_i] - \sum_{j=1}^m \frac{\mathrm{Cov}[X_i, \mathrm{Res}(X^i_j, R_j)]^2}{\mathrm{Var}[\mathrm{Res}(X^i_j, R_j)]} = \mathrm{Var}[X_i] - \sum_{j=1}^m \frac{\mathrm{Cov}[X_i, X^i_j]^2}{\mathrm{Var}[\mathrm{Res}(X^i_j, R_j)]}$.

Using this identity, and the observation about connectivity, we derive a recurrence relation for the optimum solution:

$$P(X_i, k) = \mathrm{Var}[X_i] - \max_{k_1, \ldots k_p : k_1 + \ldots k_p = k} \Big(\sum_{j=1}^{p} \frac{\mathrm{Cov}[X_i, X^i_j]^2}{P(X^i_j, k_i)} \Big).$$

We see that $P(X_i, k)$ is thus monotonically increasing in each of the $P(X^i_j, k_i)$, and in particular, optimal solutions for X_i must contain optimal solutions for the children X^i_j. This enables a bottom-up dynamic programming solution.

Because the algorithm thus needs to compute a corresponding value $P(X_i, k, j)$ for each triple, and each such computation takes time $O(k)$, the overall running time is $O(k^2 nd)$. ■

Dynamic programs for tree graphs can frequently be extended to graphs of bounded treewidth, and one would hope for a straightforward extension of the above algorithm. However, obtaining such an extension appears difficult. The reason is that if a small subset T of nodes separates the vertices into two or more partitions, and some of the nodes in T are selected, they introduce conditional dependencies between otherwise uncorrelated variables. Hence, if the subset selection problem can be solved for bounded treewidth graphs, the techniques would likely have to be significantly different.

6. COVARIANCE GRAPHS WITH LARGE INDEPENDENT SETS

Another special case amenable to polynomial time algorithms is when the covariance graph $G(C, \mathbf{b})$ has a known large independent set I, of size $|I| = n - \nu$, for some constant ν. This case is motivated by scenarios in which the algorithm has access to $n - \nu$ (known) independent samples $X_1, \ldots, X_p$, as well as ν "aggregate statistics" $Y_1, \ldots, Y_\nu$ of those variables, which may be correlated arbitrarily with the independent variables and each other. The algorithm can choose a mix of aggregate statistics Y_j and X_i variables. We present a polynomial-time algorithm for this special case.

THEOREM 6.1. *If $G(C, \mathbf{b})$ contains a (known) independent set I of size $n-\nu$, then there exists an $O(2^\nu n^{4(\nu^2+\nu+1)})$ time algorithm for the subset selection problem.*

As before, this result can be combined with the matrix perturbation bound from Lemma 3.1 to yield approximation bounds similar to those of Corollary 4.2 when $G_\epsilon(C, \mathbf{b})$ has a large independent set (assuming C is well-conditioned).

Proof. Assume that $X_1, \ldots, X_{n-\nu}$ are pairwise uncorrelated (i.e., form the independent set), while $X_{n-\nu+1}, \ldots, X_n$ may have arbitrary correlations with each other or with $X_1, \ldots, X_{n-\nu}$. Our algorithm will perform exhaustive search over all at most 2^ν subsets $T \subseteq \{X_{n-\nu+1}, \ldots, X_n\}$ of size at most $t = |T| \leq \min(k, \nu)$. For each such set T, it determines the optimum complementary set $S \subseteq \{X_1, \ldots, X_{n-\nu}\}$ of size $|S| = k - t$ in polynomial time, as described below. It then simply outputs the best set $S \cup T$ as the solution.

Given a choice of T, the objective for the remaining set S of size $k-t$ is to maximize $f(S) := R^2_{Z,S\cup T}$. We first rewrite $R^2_{Z,S\cup T}$ as follows: $R^2_{Z,S\cup T} = R^2_{Z,S} + R^2_{Z,\{\mathrm{Res}(X_j,S)|j\in T\}} = \sum_{r\in S} b_i^2 + R^2_{Z,\{X_j - \sum_{r\in S} c_{j,r}X_r | j\in T\}}$.

By Section 2, $R^2_{Z,\{X_j - \sum_{r\in S} c_{j,r}X_r|j\in T\}} = \mathbf{b}'C'^{-1}\mathbf{b}'$, where $\mathbf{b}'$ is the vector of covariances between $\mathrm{Res}(X_j, S)$ and Z for $X_j \in T$, and C' the matrix of covariances between the $\mathrm{Res}(X_j, S)$ for $X_j \in T$. Thus, their respective entries are $b'_j = b_j - \sum_{r\in S} c_{j,r}b_r$, and $c'_{i,j} = c_{i,j} - \sum_{r\in S} c_{i,r}c_{j,r}$.

Because C' is a covariance matrix, and thus positive semidefinite, the function $\mathbf{b}'C'^{-1}\mathbf{b}'$ is convex in each entry of $\mathbf{b}'$ and C' (see, e.g., Section 3.1.7 of [2]). The equivalent formulation above allows us to reduce the problem to the *Shaped Partition Problem* [17, 25], defined as follows.

A *p-shape* of n is a tuple $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ of non-negative integers such that $\sum_{i=1}^p \lambda_i = n$. Given a set Λ of p-shapes, a *p-partition* π of $\{1, \ldots, n\}$ into p disjoint sets $\pi_1, \ldots, \pi_p$ is a Λ-*partition* iff there exists a $\lambda \in \Lambda$ with $|\pi_i| = \lambda_i$ for all i. Given a set of n vectors $\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_n}$ in d dimensions and a p-partition π of $\{1, \ldots, n\}$, we let $\mathbf{B}_{\pi,\mathbf{j}} = \sum_{i\in\pi_j} \mathbf{A_i}$ be the sums of vectors in the jth partition. Given a non-empty set Λ of shapes and a convex function $g : (\mathbb{R}^d)^p \to \mathbb{R}$ defined on p-tuples of d-dimensional vectors, the shaped partition problem consists in finding a Λ-partition π maximizing $g(\mathbf{B}_{\pi,\mathbf{1}}, \mathbf{B}_{\pi,\mathbf{2}}, \ldots, \mathbf{B}_{\pi,\mathbf{p}})$.

THEOREM 6.2 (THEOREM 4.2 FROM [17]). *If the set Λ allows membership queries in time $O(1)$, the shaped partition problem is solvable in time $O(n^{dp^2})$.*

We reduce our problem to the shaped partition problem with $d = t^2 + t + 1$ and $p = 2$ as follows: For each $r \leq n - \nu$, we define a $(t^2 + t + 1)$-dimensional vector

$$\mathbf{A_r} = \left([\tfrac{b'_j}{k-t} - c_{j,r}b_r]_{j\in T},\ [\tfrac{c_{i,j}}{k-t} - c_{i,r}c_{j,r}]_{i,j\in T},\ b_r^2\right).$$

Notice that the $\mathbf{A_r}$ are explicitly defined in such a way that if S is a set of $k - t$ indices $r \leq n - \nu$, then

$$\sum_{r\in S} \mathbf{A_r} = ([b'_j - \sum_{r\in S} c_{j,r}b_r]_{j\in T},\ [c_{i,j} - \sum_{r\in S} c_{i,r}c_{j,r}]_{i,j\in T},\ \sum_{r\in S} b_r^2),$$

i.e., the entries are the entries of $\mathbf{b}'$ and C'. Hence, we can express $f(S)$ as a function of two vectors $g(\mathbf{x}, \mathbf{x}') = x_{t^2+t+1} + \mathbf{b}^{(\mathbf{x})}C^{(\mathbf{x})^{-1}}\mathbf{b}^{(\mathbf{x})}$, where $\mathbf{b}^{(\mathbf{x})}$ is the vector of coordinates $1, \ldots, t$ of $\mathbf{x}$, and $C^{(\mathbf{x})}$ the $t \times t$ matrix whose entries are the coordinates $t+1, \ldots, t+t^2$ of $\mathbf{x}$. (Notice that g does not depend on $\mathbf{x}'$.) Defining Λ as the set of all 2-partitions π of $\{1, \ldots, n-\nu\}$ with $|\pi_1| = k - t$, the sets π_1 correspond exactly to the desired sets S, and $f(S) = g(\mathbf{x}, \mathbf{x}')$.

It follows that g is convex in all its variables, and we can use the algorithm from Theorem 6.2 to determine the optimum set S in time $O(n^{4(t^2+t+1)})$. ■

7. EXPONENTIAL DECAY

In this section, we consider another important class of covariance structures: the case where the variables X_i are associated with points $y_1 \leq y_2 \leq \ldots \leq y_n$ on a line, and the covariances are $\mathrm{Cov}[X_i, X_j] = a^{|y_i - y_j|}$ for some constant $a \in (0, 1)$. This special case arises naturally in temporal processes of the following form: $X_1 = Y_1, X_i = \gamma_i X_{i-1} + Y_i$, where the Y_i are independent random variables such that each X_i has variance 1. Such processes arise naturally when a constant fraction of a population is depleted over a time period, and then refilled randomly. Since $\mathrm{Cov}[X_i, X_j] = \prod_{r=i+1}^{j} \gamma_r$ for $i < j$, this directly leads to an embedding of the X_i into the line, by setting $y_1 = 0$ and $y_i = \sum_{j=2}^{i} \log_a \gamma_j$ for $i \geq 2$, and an arbitrary constant $a \in (0, 1)$. This case can also be considered a first step toward an exploration of covariances defined by monotone decreasing functions of more general classes of metrics, a regime which is very relevant for sensor networks.

Our algorithm is based on a characterization of the inverse of such exponential-decay covariance matrices, shown in the following lemma (the proof is omitted due to lack of space).

LEMMA 7.1. *Let C be a $k \times k$ covariance matrix whose (i, j) entry is the covariance of X_{p_i} and X_{p_j}, i.e., $a^{|y_{p_i} - y_{p_j}|}$ for all i, j. We write $D_i = |y_{p_{i+1}} - y_{p_i}|$ for $1 \leq i \leq k-1$, and $D_i = \infty$ for $i = 0$ or $i = k$. With c_{ij}^{-1} denoting the entries of C^{-1}, we have*

$$c_{ij}^{-1} = \begin{cases} 0, & \text{if } |i-j| > 1 \\ -1 + \frac{1}{1-a^{2D_{i-1}}} + \frac{1}{1-a^{2D_i}}, & \text{if } j = i \\ -\frac{a^{D_i}}{1-a^{2D_i}}, & \text{if } j = i+1. \end{cases}$$

Using the above lemma, the reduction in error due to sampling a set $S = \{X_{p_1}, \ldots, X_{p_k}\}$ of k random variables is

$$\begin{aligned} \mathbf{b_S}^T C_S^{-1} \mathbf{b_S} &= \textstyle\sum_{i=1}^k \sum_{j=1}^k b_{p_i} b_{p_j} c_{i,j}^{-1} \\ &= \textstyle\sum_{i=1}^k b_{p_i}^2 (1 + \frac{a^{2D_{i-1}}}{1-a^{2D_{i-1}}} + \frac{a^{2D_i}}{1-a^{2D_i}}) - \\ &\quad 2\textstyle\sum_{i=1}^{k-1} b_{p_i} b_{p_{i+1}} \frac{a^{D_i}}{1-a^{2D_i}} \\ &= \textstyle\sum_{i=1}^k b_{p_i}^2 + \sum_{i=2}^k b_{p_i}^2 \frac{a^{2D_{i-1}}}{1-a^{2D_{i-1}}} + \\ &\quad \textstyle\sum_{i=1}^{k-1} b_{p_i}^2 \frac{a^{2D_i}}{1-a^{2D_i}} - 2\sum_{i=1}^{k-1} b_{p_i} b_{p_{i+1}} \frac{a^{D_i}}{1-a^{2D_i}} \\ &= \textstyle\sum_{i=1}^k b_{p_i}^2 + \sum_{i=1}^{k-1} (b_{p_i} - b_{p_{i+1}})^2 \frac{a^{2D_i}}{1-a^{2D_i}} - \\ &\quad 2\textstyle\sum_{i=1}^{k-1} b_{p_i} b_{p_{i+1}} \frac{a^{D_i}}{1+a^{D_i}}. \end{aligned}$$

The above equation can be used to derive a dynamic program for finding the best set S. Let $E(v, j)$ denote the maximum error reduction possible by choosing v variables, including necessarily X_j, from among variables $X_j, \ldots, X_n$. Then we obtain the recurrence relation (with $E(0, j) = 0$):

$$\begin{aligned} E(v+1, j) &= \max_{j+1 \leq i \leq n} (E(v, i) + b_j^2 + (b_j - b_i)^2 \tfrac{a^{2|y_i - y_j|}}{1-a^{2|y_i - y_j|}} \\ &\quad -2b_j b_i \tfrac{a^{|y_i - y_j|}}{1+a^{|y_i - y_j|}}). \end{aligned}$$

Thus, using dynamic programming, we can compute all the $E(v, j)$ values in time $O(n^2 k)$. The optimal error reduction using k variables is then $R^2_{Z,S^*} = \max_{1\leq j\leq n} E(k, j)$, and the actual solution can be obtained easily.

8. ABSENCE OF SUPPRESSORS

In this section, we derive another condition on covariance matrices that permits a good approximation for maximizing the error reduction $R^2_{Z,S}$. In the statistics community, so-called *suppressor variables* [5, 36] have frequently been considered as "unfavorable attributes of regression models" [38]. Intuitively, a variable X_j is a suppressor variable if it "suppresses" the correlation between some other X_i and the predictor variable Z, in the sense that X_i appears not (or only slightly) correlated with Z, but is much more correlated with Z once X_j has been sampled. For instance, if X_i and Z are independent, and $X_j = X_i + Z$, then X_j would be a suppressor variable. Formally, X_j is a suppressor variable if $|\mathrm{Corr}(Z, X_i/X_j)| > |\rho(Z, X_i)|$ for some variable X_i.

We show that a somewhat stricter version of the absence of suppressor variables leads to a performance guarantee for Forward Regression. Given a set S of random variables, we say that X_j is a *suppressor conditioned on* S iff

$$|\mathrm{Corr}(Z, \mathrm{Res}(X_i, S)/\mathrm{Res}(X_j, S))| > |\rho(Z, \mathrm{Res}(X_i, S))|.$$

Thus, the traditional notion of a suppressor variable is a suppressor conditioned on the empty set.

We then prove that in the absence of (conditional) suppressor variables, the greedy Forward Regression heuristic gives a $(1 - \frac{1}{e})$ approximation algorithm. The idea of the proof is to show that the absence of suppressor variables implies that the objective function $R^2_{Z,S}$ is submodular in S. Then, by a well-known result of Nemhauser, Wolsey, and Fisher [6, 24], the greedy algorithm for maximization is a $(1 - \frac{1}{e})$ approximation. We note here that an alternate NP-hardness proof (omitted here due to lack of space) shows that the subset selection problem remains NP-hard even in the absence of suppressors.

THEOREM 8.1. *If there are no suppressor variables conditioned on any set S, then the Forward Regression heuristic is a $(1 - \frac{1}{e})$ approximation for the problem of maximizing the error reduction, i.e., its selected set S satisfies $R^2_{Z,S} \geq (1 - \frac{1}{e})R^2_{Z,S^*}$, where S^* is the optimal k-subset.*

Proof. Recall that a function f from sets to real values is called *submodular* iff it satisfies the "diminishing returns" property $f(S+x) - f(S) \geq f(T+x) - f(T)$ whenever $S \subseteq T$. It is well-known [6, 24] that if a function is monotone, submodular, and non-negative, then the natural greedy algorithm for maximization over subsets of size k is a $(1 - 1/\mathrm{e})$ approximation. We will apply this result to the function $R^2_{Z,S}$. Obviously, $R^2_{Z,S}$ is monotone and non-negative in S, so it remains to show that it is submodular.

Consider a set $S = \{X_1, \ldots, X_n\}$, and $T = S \cup S'$, with $S' = \{Y_1, \ldots, Y_m\}$, where $m \geq 1$. Let $X \notin S \cup S'$ be another variable. We will show that $R^2_{Z,S\cup\{X\}} - R^2_{Z,S} \geq R^2_{Z,S\cup S'\cup\{X\}} - R^2_{Z,S\cup S'}$.

We let $W = \mathrm{Res}(Z, S \cup S' \cup \{X\})$, $Z' = Z - W$, and $Q = \mathrm{Res}(X, S \cup S')$, $X' = X - Q$. Thus, $Z' = \gamma X + \sum_i \alpha_i X_i + \sum_j \beta_j Y_j$, for some $\gamma, \alpha_i, \beta_j$ and $X' = \sum_i \xi_i X_i + \sum_j \eta_j Y_j$, for some ξ_i, η_j. Substituting $X = X' + Q$ into the expression for $Z = Z' + W$, and using that $\mathrm{Cov}[W, Q] = 0$, we get $Z = \sum_i(\gamma\xi_i + \alpha_i)X_i + \sum_j(\gamma\eta_j + \beta_j)Y_j + \gamma Q + W$, so $\mathrm{Res}(Z, S \cup S') = \gamma Q + W$.

Substituting the definition of squared multiple correlation, $R^2_{Z,S\cup S'\cup\{X\}} - R^2_{Z,S\cup S'} = \frac{\mathrm{Var}[Z'] - (\mathrm{Var}[Z] - \gamma^2\mathrm{Var}[Q] - \mathrm{Var}[W])}{\mathrm{Var}[Z]} = \gamma^2\mathrm{Var}[Q]$.

Also, the definition of semi-partial correlation and of Q gives that $\mathrm{Corr}(Z, \mathrm{Res}(X, S\cup S')) = \frac{\mathrm{Cov}[Z,\mathrm{Res}(X,S\cup S')]}{\sqrt{\mathrm{Var}[Z]}\sqrt{\mathrm{Var}[\mathrm{Res}(X,S\cup S'Y)]}}$ $= \frac{\mathrm{Cov}[Z'+W,Q]}{\sqrt{\mathrm{Var}[Q]}} = \frac{\mathrm{Cov}[Z',Q]}{\sqrt{\mathrm{Var}[Q]}} = \gamma\sqrt{\mathrm{Var}[Q]}$.

Hence, $R^2_{Z,S\cup S'\cup\{X\}} - R^2_{Z,S\cup S'} = \mathrm{Corr}(Z, \mathrm{Res}(X, S\cup S'))^2$. An identical proof gives $R^2_{Z,S\cup\{X\}} - R^2_{Z,S} = \mathrm{Corr}(Z, \mathrm{Res}(X, S))^2$. But by the original assumption on the absence of conditional suppressors, we have that for any $Y_i \in S'$, $|\mathrm{Corr}(Z, \mathrm{Res}(X, S\cup\{Y_i\}))| = |\mathrm{Corr}(Z, \mathrm{Res}(X, S)/\mathrm{Res}(Y_i, S))| \leq |\rho(Z, \mathrm{Res}(X, S))|$, from which it can be seen by a simple inductive proof that $|\mathrm{Corr}(Z, \mathrm{Res}(X, S\cup S'))| \leq |\rho(Z, \mathrm{Res}(X, S))|$. Taking squares now completes the proof. ■

9. CONCLUSIONS AND OPEN PROBLEMS

We investigated the problem of selecting a subset S of observation variables for linear regression to maximize $R^2_{Z,S}$. We gave exact algorithms for several restricted classes of covariance structures including the case where $G(C, \mathbf{b})$ forms a tree, or has a large independent set, and an FPTAS when $\tilde{G}(C)$ forms a constant-bandwidth graph. Using our perturbation results, we could then extend these results for the cases when the graphs violated these structures by edges with small covariances. We also gave exact and approximation results for certain exponentially decaying covariance matrices and for cases without suppressor variables.

Naturally, the most important direction for future work is to obtain approximation guarantees for other, more general, cases. While the hardness result mentioned in the text precludes any general approximation for minimizing the mean square prediction error, no hardness result is known for the problem of maximizing the error reduction. The difficulty in either a hardness or general approximation result lies in the highly non-linear behavior of the matrix inverse.

Barring a more general approximation or hardness result, one direction for future work is to identify interesting and practically relevant special cases for which tractable algorithms can be obtained. One could hope that the results from Section 5 would extend to bounded treewidth; however, such an extension appears not as natural as in many other cases of dynamic programming algorithms for trees.

The results of Section 7 suggest extensions of our work in at least two directions. The motivation from temporal processes suggests considering a higher-order Markov process, of the form $X_{i+1} = \sum_{j=i-p}^{i} \gamma_{i+1,j} X_j + Y_{i+1}$. The motivation from sensor networks would aim at generalizing the types of dependencies on the distance beyond exponential, and the metric space beyond one-dimensional.

Beyond the prediction of a single variable Z, it would be of interest to predict multiple variables $Z_1, \ldots, Z_r$ at once, with an appropriately chosen aggregation measure of prediction quality. We leave these questions as interesting directions for future work.

Acknowledgments

We would like to thank Muthu Muthukrishnan and several anonymous referees for useful feedback.

10. REFERENCES

[1] K. Anstreicher, M. Fampa, J. Lee, and J. Williams. Maximum-entropy remote sampling. *Discrete Applied Mathematics*, 108(3):211–226, 2001.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

[4] W. Cochran. Some effects of errors of measurement on multiple correlation. *Journal of the American Statistical Association*, 65(329):22–34, 1970.

[5] J. Cohen and P. Cohen. *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erlbaum Assoc Publishers, 2003.

[6] G. Cornuejols, M. Fisher, and G. Nemhauser. Location of bank accounts to optimize float. *Management Science*, 23:789–810, 1977.

[7] C. Couvreur and Y. Bressler. On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM Journal on Matrix Analysis and Applications*, 21(3):797–808, 2000.

[8] G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. *Journal of Constructive Approximation*, 13:57–98, 1997.

[9] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong. Model driven data acquisition in sensor networks. In *Proc. International Conference on Very Large Data Bases, VLDB*, 2004.

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

[11] D. Donoho. For most large underdetermined systems of linear equations, the minimal 11-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2005.

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

[13] V. F. Flack and P. C. Chang. Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician Journal*, 41(1):84–86, 1987.

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

[15] C. Guestrin, A. Krause, and A. Singh. Near-optimal sensor placements in gaussian processes. In *International Conference on Machine Learning, ICML*, 2005.

[16] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1999.

[17] F. Hwang, S. Onn, and U. Rothblum. A polynomial time algorithm for shaped partition problems. *SIAM Journal on Optimization*, 10(1):70–81, 1999.

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

[19] C.-W. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

[20] P. Liaskovitis and C. Schurgers. Leveraging redundancy in sampling-interpolation applications for sensor networks. In *Proc. 3rd Intl. Conf. on Distributed Computing in Sensor Systems*, 2007.

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

[22] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1, 2005.

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

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

[25] S. Onn and L. Schulman. The vector partition problem for convex optimization functions. *Mathematics of Operations Research*, 26(3):583–590, 2001.

[26] M. H. Pesaran and R. J. Smith. A generalized r2 criterion for regression models estimated by the instrumental variables method. *Econometrica*, 62(3):705–710, 1994.

[27] J. Saxe. Dynamic programming algorithms for recognizing small bandwidth graphs in polynomial time. *SIAM Journal on Algebraic Methods I*, 1(4):363–369, 1980.

[28] G. W. Stewart and J.G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[29] V. Temlyakov. Greedy algorithms and m-term approximation with regard to redundant dictionaries. *Journal of Approximation Theory*, 98:117–145, 1999.

[30] V. Temlyakov. Nonlinear methods of approximation. *Foundations of Computational Mathematics*, 3:33–107, 2002.

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

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

[33] J. Tropp. *Topics in Sparse Approximation*. PhD thesis, University of Texas, Austin, 2004.

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

[35] J. Tropp, A. Gilbert, S. Muthukrishnan, and M. Strauss. Improved sparse approximation over quasi-incoherent dictionaries. In *Proc. IEEE-ICIP*, 2003.

[36] W. F. Velicer. Suppressor variables and the semipartial correlation coefficient. *Educational and Psychological Measurement*, 38:953–958, 1978.

[37] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using l1-constrained quadratic programming. In *Proc. Allerton Conference on Communication*, 2006.

[38] D. A. Walker. Suppressor variable(s) importance within a regression model. *Journal of College Student Development*, 44:127–133, 2003.

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