Variable Inclusion and Shrinkage Algorithms

PETER RADCHENKO AND GARETH M. JAMES *

Abstract

The Lasso is a popular and computationally efficient procedure for automatically performing both variable selection and coefficient shrinkage on linear regression models. One limitation of the Lasso is that the same tuning parameter is used for both variable selection and shrinkage. As a result, it typically ends up selecting a model with too many variables to prevent over shrinkage of the regression coefficients. We suggest an improved class of methods called "Variable Inclusion and Shrinkage Algorithms" (VISA). Our approach is capable of selecting sparse models while avoiding over shrinkage problems and uses a path algorithm so is also computationally efficient. We show through extensive simulations that VISA significantly outperforms the Lasso and also provides improvements over more recent procedures, such as the Dantzig selector, Relaxed Lasso and Adaptive Lasso. In addition, we provide theoretical justification for VISA in terms of non-asymptotic bounds on the estimation error that suggest it should exhibit good performance even for large numbers of predictors. Finally, we extend the VISA methodology, path algorithm, and theoretical bounds to the Generalized Linear Models framework.

Some key words: Variable Selection; Lasso; Generalized Linear Models; Dantzig Selector.

1 Introduction

When fitting the traditional linear regression model,

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2), \quad i = 1,\ldots n, \tag{1}$$

with the number of predictors, p, large relative to the sample size, n, there are many approaches that outperform ordinary least squares (OLS) (Frank and Friedman, 1993). Most of the alternatives can be categorized into one of two groups. The first set of approaches uses some form of regularization on the regression coefficients to trade off increased bias

*Marshall School of Business, University of Southern California. This work was partially supported by NSF Grant DMS-0705312.

for a possibly significant decrease in variance. While these approaches often produce improvements in prediction accuracy, the final fit may be difficult to interpret because all p variables will remain in the model. The second set of approaches begins by performing variable selection i.e. determining which $\beta_j \neq 0$. By implementing OLS on the reduced number of variables one can often gain both increased prediction accuracy as well as a more easily interpretable model.

More recently interest has focused on an alternative class of methods which implement both the variable selection and the coefficient shrinkage in a single procedure. The most well known of these procedures is the Lasso (Tibshirani, 1996; Chen *et al.*, 1998). The Lasso uses an L_1 penalty on the coefficients, which has the effect of automatically performing variable selection by setting certain coefficients to zero and shrinking the remainder. This method was made particularly appealing by the advent of the LARS algorithm (Efron *et al.*, 2004) which provided a highly efficient means to simultaneously produce the set of Lasso fits for all values of the tuning parameter. Numerous improvements have been suggested for the Lasso. A few examples include the adaptive Lasso (Zou, 2006), SCAD (Fan and Li, 2001), the Elastic Net (Zou and Hastie, 2005), CAP (Zhao *et al.*, 2006), the Dantzig selector (Candes and Tao, 2007), the Relaxed Lasso (Meinshausen, 2007), and the Double Dantzig (James and Radchenko, 2008).

The main limitation of the Lasso is that in situations where the true number of non-zero coefficients is small relative to p, it must choose between including a number of irrelevant variables or else over shrinking the correct variables in order to produce a model of the correct size. This tradeoff is caused by the fact that the Lasso uses a single tuning parameter to control both the variable selection and the shrinkage component of the fitting procedure. The Relaxed Lasso attacks this problem directly by introducing a second tuning parameter. The first parameter controls the number of variables that are included in the model while the second controls the level of shrinkage on the selected variables. Meinshausen (2007) shows that the Relaxed Lasso can significantly outperform the Lasso when p is large relative to n and provides comparable performance to the Lasso in other situations.

In this paper we suggest a new approach called the "Variable Inclusion and Shrinkage Algorithm" (VISA). As with the Relaxed Lasso we utilize two tuning parameters. However, rather than using the hard thresholding approach enforced by the Relaxed Lasso and other variable selection methods, where only variables included according to the first tuning parameter may enter the model, we allow for the potential inclusion of all variables. Our first tuning parameter divides the variables into two groups. The first group receives preference for model inclusion but variables from the second group may still be included if there is evidence that they are significant. Thus errors in the original variables can be eliminated to produce the correct model.

This paper makes four key contributions. First, we demonstrate through simulations and real world examples that VISA consistently produces considerable improvements over the Lasso, as well as related approaches, such as the Dantzig Selector and the adaptive Lasso. It also provides statistically significant improvements over the Relaxed Lasso. Second, we develop a fitting algorithm, similar in nature to LARS, for efficiently computing the entire

sample path of VISA coefficients. Hence, the computation cost is similar to that for the Lasso or Relaxed Lasso, both of which are considered to be extremely efficient procedures. Third, we provide theoretical results demonstrating situations where both the Lasso and Relaxed Lasso will fail but VISA will generate the correct model. In addition we show that VISA possesses similar types of non-asymptotic bounds to those that Candes and Tao (2007) proved for the Dantzig selector. The Dantzig selector's non-asymptotic bounds have attracted a great deal of attention because they show that the L_2 error in the estimated coefficients is within a factor of $\log p$ of that one could achieve if the true model were known. These bounds suggest VISA should have good levels of performance even for p much larger than n. Finally, we extend both the VISA fitting algorithms and the theoretical bounds to the more general class of Generalized Linear Models (McCullagh and Nelder, 1989). To our knowledge, this is the first time that bounds of this form have been proposed for GLM's.

The remainder of this paper is organized as follows. In Section 2 we explicitly define the VISA approach for linear models and develop an efficient path fitting procedure. Our non-asymptotic bounds and other theoretical contributions are also provided in this section. A comprehensive simulation comparison of VISA with the Lasso, Dantzig selector, Relaxed Lasso and other methods is presented in Section 3. We demonstrate the practical performance of VISA on two real world data sets in Section 4. Finally, Section 5 extends the VISA methodology and theory to GLM's and Section 6 provides a discussion.

2 Methodology

In this section we describe the VISA methodology. VISA is in fact a general class of approaches. In Section 2.1 we explain the general VISA implementation. Then in Sections 2.2 and 2.3 we provide two specific implementations of this approach using modified versions of the LARS (Efron *et al.*, 2004) and DASSO (James *et al.*, 2008) algorithms. Our theoretical results are presented in Section 2.4.

2.1 General VISA Methodology

Using suitable location and scale transformations we can standardize the predictors and center the response, so that

$$\sum_{i=1}^{n} Y_i = 0, \quad \sum_{i=1}^{n} X_{ij} = 0, \quad \sum_{i=1}^{n} X_{ij}^2 = 1, \quad \text{for } j = 1, \ldots, p. \tag{2}$$

Throughout the paper we assume that (2) holds. However, all numerical results are presented on the original scale of the data.

Figure 1 illustrates the differences between the Lasso, Dantzig selector, Relaxed Lasso and our VISA methodology. Let $c_j = \mathbf{x}_j^T(\mathbf{Y} - X\hat{\beta})$ denote the covariance between the jth predictor and the residual vector. The solid line in the first plot of Figure 1 represents the

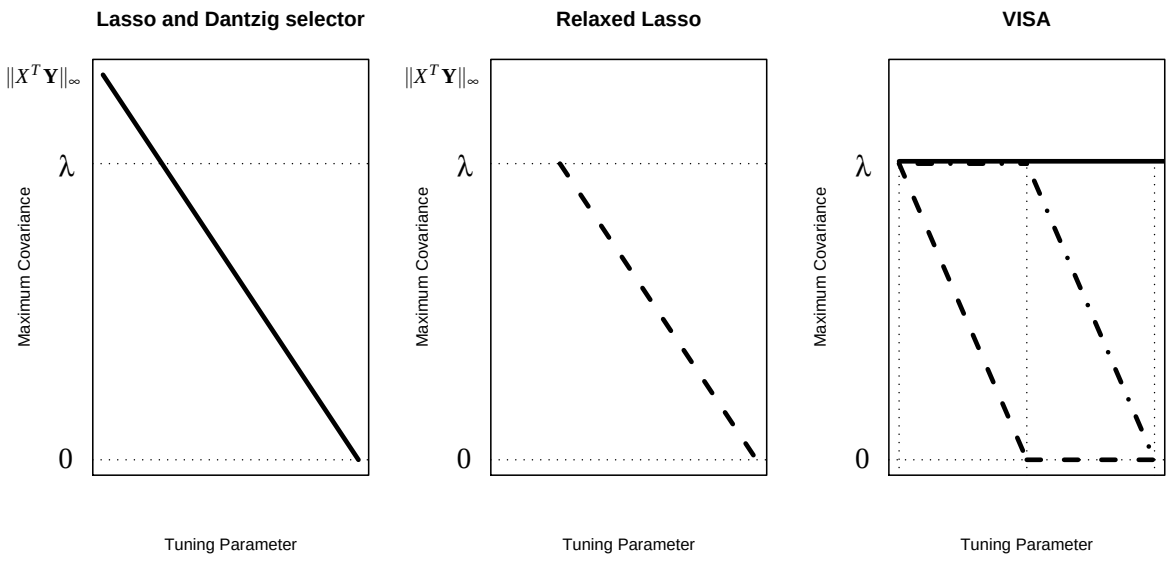

Figure 1: *Here the solid, dashed and dash-dot lines respectively represent constraints applied to all the variables, a primary subset of variables, and a secondary subset of variables.*

maximum absolute c_j among all predictors, i.e. $\|X^T(\mathbf{Y} - X\hat{\beta})\|_\infty$, as a function of the tuning parameter. The Lasso and Dantzig selector both construct coefficient paths beginning with the maximum covariance equal to $\|X^T\mathbf{Y}\|_\infty$ and systematically reducing it to zero. One then chooses a tuning parameter λ and selects the point on the path where the maximum covariance equals λ. We denote this point by $\beta_\lambda(0)$. It has been well documented (Meinshausen, 2007; Candes and Tao, 2007; James and Radchenko, 2008) that these methods must either select a large λ which results in a sparse model but over shrinkage of the coefficients or else a low value of λ which reduces the shrinkage problem but also tends to introduce many undesired variables into the model.

The Relaxed Lasso (Meinshausen, 2007), illustrated in the second plot of Figure 1, proposes a solution to this problem. First, for a given value of λ, the Lasso is used to find the point $\beta_\lambda(0)$ where the maximum covariance equals λ. The variables at the covariance boundary $\mathcal{A}_1 = \{j : |c_j| = \lambda\}$ are selected for the final model. The Relaxed Lasso then continues the path by driving the covariances for $\mathcal{A}_1$ to zero, as represented by the dashed line. This approach allows one to select a sparse model without requiring over shrinkage on the nonzero coefficients. However, any errors in the initial model selected by $\beta_\lambda(0)$ can not be corrected.

Our VISA methodology, illustrated in the final plot of Figure 1, also removes the over shrinkage problem, but without the requirement to permanently exclude variables. As with the Relaxed Lasso, a value for λ is chosen, $\beta_\lambda(0)$ is computed as the starting point for the coefficient path, we identify the variables $\mathcal{A}_1$, and drive their covariances to zero. We call $\mathcal{A}_1$ the "primary" variables. However, VISA differs from the Relaxed Lasso in several key respects. First, at no point do we fix any of the coefficients to zero and hence, as the fit to

our data improves, new variables may always enter the model. As a result, mistakes in the primary variables can be corrected. Second, while driving the covariances towards zero we require that at all points of the coefficient path

$$\|X^T(\mathbf{Y} - X\hat{\beta})\|_\infty \leq \lambda. \tag{3}$$

One can show that, for appropriate λ, bound (3) will hold with high probability for the true coefficient vector, β, thus it is sensible to enforce this constraint on the estimated coefficients. In addition, (3) provides an automatic form of variable selection, because if a covariance reaches the threshold $\pm\lambda$ as the primary covariances are driven to zero, a new predictor will be added to the model to maintain the constraint. Finally, as represented by the dash-dot line, after the covariances for the primary variables have reached zero, we then identify the "secondary" set of variables currently at the boundary $|c_j| = \lambda$, and send their covariances to zero.

We provide theoretical justification for the VISA method in Section 2.4, and illustrate its excellent practical performance via a comprehensive simulation study in Section 3. However, the intuition for this three step approach can be described as follows, using the VISA$_L$ implementation introduced in the next subsection. We fix the value of λ, and in the first step of our algorithm select a sparse initial Lasso estimate $\beta_\lambda(0)$ that satisfies (3). In the second step we use the c_j's as measures of the "current" importance of each variable. The variables with maximum importance are selected as the "primary" predictors, and a path $\beta_\lambda(s)$ is constructed that systematically drives their $|c_j|$'s towards zero, while maintaining (3) for the remaining predictors. The primary variables represent our initial guess for the model, and by contracting their covariances we reduce or eliminate over shrinkage on their coefficients. However, while the focus is on the primary variables, at all stages of our algorithm any predictor may enter the model. For instance, we show in Section 2.4 that, under certain conditions, the "true" variables, i.e. those related to Y, that miss being classified as primary variables may see their $|c_j|$'s rise until they reach the threshold λ. At this point these predictors will be included in the model, providing an automatic correction for initial "mistakes". This self correcting property is not possible with other approaches such as the Relaxed Lasso. Finally, once the primary covariances have reached zero, in the third step of our algorithm we identify a secondary set of variables whose $|c_j|$'s are now at the boundary λ. This secondary set of predictors can be thought of as newly identified model variables. Because they have large covariances, it is likely that their coefficients will also be over shrunk, so the final section of the VISA path involves driving the secondary $|c_j|$'s towards zero, while maintaining the primary covariances at zero.

2.2 VISA Using LARS

Here we detail the algorithm for implementing one version of our VISA methodology. We call this approach VISA$_L$ because it involves an adaptation of the LARS algorithm. Throughout the algorithm, index set $\mathcal{A}_\downarrow$ represents the variables whose absolute covariances

are being simultaneously driven to zero, and C denotes the common value of all those covariances. Index set $\mathcal{A}_=$ represents the variables whose covariances are being held constant, either at the levels $\pm\lambda$ or at zero. Write $\mathcal{A}$ for $\mathcal{A}_\downarrow \cup \mathcal{A}_=$ and let $X_\mathcal{A}$ be the matrix consisting of the columns of X in $\mathcal{A}$. Let $\mathbf{s}_\mathcal{A}$ be an $|\mathcal{A}|$-dimensional vector with components corresponding to $\mathcal{A}_\downarrow$ equal to the signs of the respective c_j's, and the components corresponding to $\mathcal{A}_=$ equal to zero. Finally, recall that $\beta_\lambda(0)$ is the Lasso solution for which $\max|c_j| = \lambda$. The VISA$_L$ algorithm consists of the following steps.

1. Initialize $\beta^1 = \beta_\lambda(0)$, $\mathcal{A}_\downarrow = \{j : |c_j| = \lambda\}$, $\mathcal{A}_= = \emptyset$, and $l = 1$.

2. Compute the $|\mathcal{A}|$-dimensional direction vector $\mathbf{h}_\mathcal{A} = \left(X_\mathcal{A}^T X_\mathcal{A}\right)^{-1}\mathbf{s}_\mathcal{A}$. Let $\mathbf{h}$ be the p-dimensional vector with the components corresponding to $\mathcal{A}$ given by $\mathbf{h}_\mathcal{A}$ and the remainder set to zero.

3. Compute γ, the distance to travel in direction $\mathbf{h}$ until $C = 0$ or a new $|c_i|$ reaches the level λ. Define $\beta^{l+1} = \beta^l + \gamma\mathbf{h}$, add index i to $\mathcal{A}_=$, and set $l \leftarrow l + 1$.

4. Repeat steps 2 and 3 until $C = 0$.

5. Rename $\mathcal{A}_\downarrow$ and $\mathcal{A}_=$ to $\mathcal{A}_=$ and $\mathcal{A}_\downarrow$, respectively. Repeat steps 2 and 3 until $C = 0$.

The coefficient path $\beta_\lambda(\cdot)$ is constructed by linearly interpolating $\beta^1, \beta^2, \ldots, \beta^L$, where β^L is the endpoint of the algorithm. The starting point, $\beta_\lambda(0)$, can be computed easily by implementing the first few steps of LARS. Steps 2 through 4 produce the red dashed line, from Figure 1, corresponding to the first part of the VISA path where the covariances for the primary variables are driven to zero. Step 5, corresponding to the dash-dot line, then identifies the secondary variables and sends their covariances to zero. The zeros in $\mathbf{s}_\mathcal{A}$ ensure that the $|c_j|$'s in $\mathcal{A}_=$ remain fixed at either λ or zero, while the ± 1's in $\mathbf{s}_\mathcal{A}$ ensure that the $|c_j|$'s in $\mathcal{A}_\downarrow$ decrease at a constant rate. Note that when $\mathbf{s}_\mathcal{A}$ has no zeroes, $\mathbf{h}$ is simply the LAR direction. Similarly to LAR, γ is computed using

$$\gamma = \min_{j \notin \mathcal{A}}^{+} \left\{ \frac{c_j - \lambda}{\mathbf{x}_j^T X \mathbf{h}}, \frac{c_j + \lambda}{\mathbf{x}_j^T X \mathbf{h}}, \frac{c_k}{\mathbf{x}_k^T X \mathbf{h}} \right\},$$

where the minimum is taken over the positive components, and k is some member of $\mathcal{A}_\downarrow$.

In our simulations we use a slightly modified version of the above algorithm by forcing the first computed direction to be LARS-Lasso rather than LAR. This is done by the following small change to step 1: if the LARS path that produced $\beta_\lambda(0)$ has a coefficient that hit zero at the breakpoint corresponding to λ, remove this coefficient from the set $\mathcal{A}_\downarrow$.

2.3 VISA Using DASSO

In this section we detail the algorithm for implementing another version of our VISA methodology, VISA$_D$. This approach uses a modified version of DASSO (James *et al.*,

2008). The DASSO algorithm generates the entire path of the Dantzig selector solutions. The Dantzig selector (Candes and Tao, 2007) is defined as the solution to

$$\text{minimize} \quad \|\tilde{\beta}\|_1 \quad \text{subject to} \quad \|X^T(\mathbf{Y} - X\tilde{\beta})\|_\infty \leq \lambda \tag{4}$$

but it can also be viewed as the value that minimizes $\|X^T(\mathbf{Y} - X\tilde{\beta})\|_\infty$ subject to $\|\tilde{\beta}\|_1 \leq s$. This is the same optimization as for the Lasso, except that the loss function involves the maximum of the partial derivatives of the sum of squares. James *et al.* (2008) demonstrate strong connections between the Lasso and Dantzig selector and also between DASSO and LARS. The main difference is that with LARS, when a new variable enters the active set, $\mathcal{A}$, this variable will automatically enter the model, while with DASSO this is not always the case. LARS adjusts the coefficients in the direction $\left(X_\mathcal{A}^T X_\mathcal{A}\right)^{-1}\mathbf{s}_\mathcal{A}$, where $\mathbf{s}_\mathcal{A}$ is a vector of 1's and -1's. However, while this direction systematically reduces the maximum absolute covariance, it does not always result in the largest reduction per unit increase in $\|\beta\|_1$. By comparison, for each active set, $\mathcal{A}$, DASSO computes a corresponding set, $\mathcal{B}$, indexing the variables in the model, i.e. those with non-zero coefficients. With LARS $\mathcal{A}$ and $\mathcal{B}$ are identical, but DASSO chooses $\mathcal{B}$ so that the direction $\left(X_\mathcal{A}^T X_\mathcal{B}\right)^{-1}\mathbf{s}_\mathcal{A}$ produces the greatest reduction in the maximum absolute covariance per unit increase in $\|\beta\|_1$.

DASSO can be thought of as an extension of LARS. We adapt VISA$_L$ in a similar manner to produce VISA$_D$. The set $\mathcal{B}$ is initialized to index the nonzero coefficients of $\beta_\lambda(0)$, which is now the solution to problem (4). The active set, $\mathcal{A}$, is now defined as $\mathcal{A}_\downarrow^M \cup \mathcal{A}_=$, where $\mathcal{A}_\downarrow^M$ corresponds to the variables in $\mathcal{A}_\downarrow$ with the maximum absolute covariance. Steps 2 and 3 of VISA$_L$ are replaced by

2. Identify either the index to be added to $\mathcal{B}$ or the index to be removed from $\mathcal{A}_\downarrow^M$. Use the new $\mathcal{A}$ and $\mathcal{B}$ to calculate the $|\mathcal{B}|$-dimensional direction vector $\mathbf{h}_\mathcal{B} = \left(X_\mathcal{A}^T X_\mathcal{B}\right)^{-1}\mathbf{s}_\mathcal{A}$. Let $\mathbf{h}$ be the p-dimensional vector with the components corresponding to $\mathcal{B}$ given by $\mathbf{h}_\mathcal{B}$ and the remainder set to zero.

3. Compute γ, the shortest distance to travel in direction $\mathbf{h}$ until a new $|c_j|$ reaches the level λ, a new $|c_m|$ hits C for $m \in \mathcal{A}_\downarrow$, a coefficient path crosses zero, or $C = 0$. Define $\beta^{l+1} = \beta^l + \gamma\mathbf{h}$ and add i to $\mathcal{A}_=$ or m to $\mathcal{A}_\downarrow^M$. Set $l \leftarrow l + 1$.

The details for these steps are provided in the appendix, while all other steps are unchanged from those of VISA$_L$.

A key advantage of VISA$_D$ derives from the fact that the path of coefficients it generates can be shown to provide the set of all solutions $\beta_\lambda(s)$ that minimize $\|\tilde{\beta}\|_1$ subject to

$$\begin{aligned} & \|X^T\,(Y - X\tilde{\beta})\|_\infty \;\leq\; \lambda, \qquad \text{for } s \text{ in } [0, 2\lambda] \\ & \|X_{\mathcal{A}_1}^T(Y - X\tilde{\beta})\|_\infty \;\leq\; \lambda - s, \quad \text{for } s \text{ in } (0, \lambda] \qquad (5) \\ \|X_{\mathcal{A}_1}^T(Y - X\tilde{\beta})\|_\infty = 0, \quad & \|X_{\mathcal{A}_2}^T(Y - X\tilde{\beta})\|_\infty \;\leq\; 2\lambda - s, \text{ for } s \text{ in } (\lambda, 2\lambda] \qquad (6) \end{aligned}$$

where λ and s are tuning parameters. Here we define $\mathcal{A}_1 = \{j : |c_j| = \lambda\}$ when solving the above optimization problem with $s = 0$, and $\mathcal{A}_2 = \{j : |c_j| = \lambda\}$ when solving it

with $s = \lambda$. The proof of this result is analogous to that given for the DASSO in James *et al.* (2008). Notice that for $s = 0$ the Dantzig selector and VISA$_D$ optimization criteria are identical. However, as s increases, the covariances in $\mathcal{A}_1$ are driven down to zero, and the level of shrinkage on the primary variables is reduced. For $s \geq \lambda$ the shrinkage on the primary variables has been eliminated, and the newly identified secondary variables have their shrinkage reduced, as the covariances in $\mathcal{A}_2$ are driven to zero. Equations (5) and (6) can both be formulated as linear programming problems, so, as an alternative to computing the entire path using the previously mentioned algorithm, we can also efficiently compute the VISA$_D$ solution for any given λ and s.

2.4 Theoretical Results

In this section we give some theoretical results providing justification for the VISA approach. First, in Section 2.4.1 we illustrate a scenario where one can prove that the LARS and Relaxed Lasso methods will fail but VISA will still produce the correct model. Then in Section 2.4.2 we show that VISA possesses non-asymptotic bounds on its estimation errors which suggest good performance for large p.

2.4.1 Comparison Between VISA and LARS

Lemma 1 below outlines a scenario, with two "signal" and many "noise" variables, where one can provide general conditions such that VISA will choose the correct model even in situations where LARS will fail. Recall the linear model (1) and view the model size p_n as a function of n. Suppose that β has two non-zero coefficients, $\beta = (\beta_1, \beta_2, 0, ..., 0)$, and write ρ_{lk} for the sample correlation between X_l and X_k. Note that we suppress the dependence on the sample size n to simplify the notation. Write $a_n \gg b_n$ and $c_n \gtrsim d_n$ to mean $b_n/a_n \to 0$ and $d_n/c_n = O(1)$ as n tends to infinity. Denote by $\beta^{ols}_{1,2}$ the OLS solution using a model with only X_1 and X_2 and denote by J_n the set of indexes $\{3, ..., p_n\}$ corresponding to the noise variables.

Lemma 1 *Let β_1 and β_2 be positive. Suppose that there exists a positive δ, such that for each n the correlations $\{\rho_{1j}, \rho_{2j}, j \in J_n\}$ lie in $(\delta, 1-\delta)$ and $|\rho_{12}| < 1-\delta$. Assume that $\beta_1 \gtrsim n^{1/2}$ and $n^{1/2} \gg \beta_2 \gg \sqrt{\log p_n}$.*

1. *If for some positive δ_1, the inequality $\rho_{1j} + \rho_{2j} < 1 + \rho_{12} - \delta_1$ holds for all j in J_n and all n, then, with probability tending to one, LARS can identify the correct model, and there is a VISA$_L$ coefficient path that identifies the correct model and stops at $\beta^{ols}_{1,2}$.*

2. *If for some positive δ_2, the inequality $\max_{J_n}(\rho_{1j} + \rho_{2j}) > 1 + \rho_{12} + \delta_2$ holds for all n, then, with probability tending to one, LARS cannot identify the correct model. If, in addition, there exists a positive δ_3, such that*

$$\max_{j \in J_n} \left(\frac{\rho_{2j} - \rho_{1j}\rho_{12}}{1 - \rho_{1j}} \right) < 1 - \rho_{12}^2 - \delta_3 \tag{7}$$

for all n, then, with probability tending to one, there is a VISA$_L$ coefficient path that identifies the correct model and stops at $\beta^{ols}_{1,2}$.

The assumption here that the β's grow with n is reasonable because these are the coefficients after normalizing $\mathbf{x}_j$. Hence as n grows the β's must also to maintain the original scale. Figure 2 provides a graphical representation of the regions where VISA and LARS will differ. For this figure we have assumed common correlations between signal and noise variables, ρ_{1j} and ρ_{2j}, and have plotted them on the x axis. The correlation between the two signal variables, ρ_{12}, is plotted on the y axis. In the middle region both methods work. Alternatively, in the two side regions LARS fails while VISA will still choose the correct model. Finally, when ρ_{1j} and ρ_{2j} become too large neither method works. However, it is worth noting that even in this third region VISA could be adapted to choose the correct model at the expense of introducing a third tuning parameter. We have not explored that option here because of the practical problems associated with three tuning parameters. If LARS can not identify the correct model, then the Relaxed Lasso will also fail, so Lemma 1 applies equally well to the latter method. With some additional assumptions, the results from Lemma 1 can also be extended to the LARS-Lasso, Dantzig selector and VISA$_D$. Lemma 1 refers to a simplified situation with only two signal variables but the ideas also apply to more complicated situations.

2.4.2 Non-Asymptotic Bounds on VISA Errors

As before, we assume that the columns of X have been standardized and that σ is the standard deviation of the errors, ε_i. Given an index set $J \subset \{1,\ldots,p\}$, write X_J for the n by $|J|$ submatrix obtained by extracting the columns of X corresponding to the indices in J.

Definition 1 *Let* $\phi(k)$ *denote the smallest eigenvalue of the matrices in* $\{X_J^T X_J, |J| \leq k\}$.

Note that $\phi(k)$ is positive if all subsets of k columns of X are linearly independent. Also note that $X_J^T X_J$ is a sample correlation matrix for the variables specified by the subset J. Then Theorem 1 allows us to place a non-asymptotic bound on the L_2 error in the VISA estimate.

Theorem 1 *Suppose that* $\beta \in \mathbb{R}^p$ *is an S-sparse coefficient vector. Consider an* $a > 0$, *and define* $\tau_p = \sigma\sqrt{2(1+a)\log p}$. *If* $\hat{\beta}$ *is a VISA estimator with k non-zero* $\hat{\beta}_j$ *coefficients for which* $\beta_j = 0$, *and* $\lambda_\infty = \|X^T(Y - X\hat{\beta})\|_\infty$, *then*

$$P\left(||\hat{\beta} - \beta||_2 > \frac{(S+k)^{1/2}}{\phi(S+k)}(\lambda_\infty + \tau_p)\right) \leq \left(p^a\sqrt{4\pi\log p}\right)^{-1}.$$

By construction of VISA, λ_∞ is bounded by the tuning parameter λ. In addition, in the simulation study of Section 3, for the solutions chosen using a validation set, we found λ_∞ to be always lower than τ_p. Hence the bound is generally proportional to $\sigma(S+k)^{1/2}\sqrt{2\log p}$ which has a similar form to that of the Dantzig selector. The main differences are that our

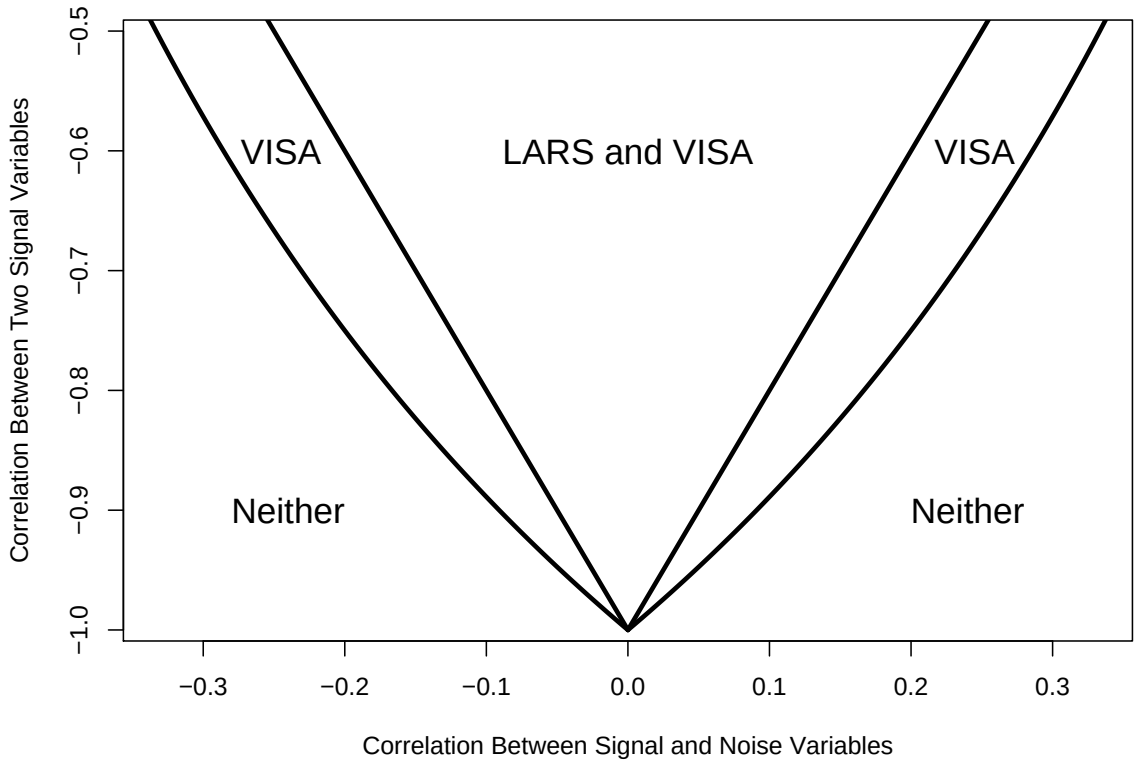

Figure 2: *Possible scenarios for different correlations between signal and noise variables (x axis) and between the two signal variables (y axis). VISA and LARS both choose the correct model in the center region. Only VISA chooses the correct model in the two side regions, and neither succeed in the lower region.*

bound requires a smaller constant but involves k, which is not present for the Dantzig selector. Of course, in practice k can always be bounded by the number of non-zero coefficients in $\hat{\beta}$.

The bound presented in Theorem 1 assumes that X has columns of norm one. If we let the sample size n grow, then maintaining the norm one columns would involve growing the individual entries of the coefficient vector β. To understand the implications of the error bounds for the parameter vector on its original scale, we now consider the situation where the columns are assumed to have squared norms proportional to n. Suppose that instead of normalizing the predictor vectors, we rescale them to make their squared norms equal $\vartheta^2 n$, where ϑ^2 is the average squared element of the original design matrix. Denote by β^* and $\hat{\beta}^*$ the corresponding rescaled versions of the true parameter vector and the VISA solution. The following result is a direct consequence of Theorem 1.

Corollary 1 *Under the assumptions of Theorem 1,*

$$P\left(||\hat{\beta}^* - \beta^*||_2 > \frac{1}{\sqrt{n}} \frac{(S+k)^{1/2}}{\vartheta\phi(S+k)} (\lambda_\infty + \tau_p)\right) \leq \left(p^a \sqrt{4\pi \log p}\right)^{-1}. \tag{8}$$

Corollary 1 makes it easier to trace the estimation error of the VISA estimator as n and p tend to infinity. In particular, if the ratio $(S+k)^{1/2}/(\vartheta\phi(S+k))$ stays stochastically bounded and $\lambda_\infty < \tau_p$, then $||\hat{\beta}^* - \beta^*||_2 = O_p\left(n^{-1/2}\sqrt{\log p}\right)$.

3 Simulation Study

In this section we present a detailed simulation study comparing VISA$_L$ and VISA$_D$ to five competing methods. We conducted a total of 48 simulations. Tables 1 and 2 report results from a representative sampling of nine of the simulations. Each simulation compared VISA$_L$ and VISA$_D$ to the Double Dantzig (DD), Relaxed Lasso (Relaxo), Adaptive Lasso, Dantzig selector and Lasso. The Double Dantzig uses one tuning parameter to perform variable selection and a second to adjust the level of shrinkage on the selected variables, in a similar fashion to the Relaxed Lasso. The main difference is that it uses the Dantzig selector criteria rather than the Lasso's. The Adaptive Lasso uses the least squares fit to reweight the variables and then produces a Lasso fit based on the reweighted predictors. Since these approaches are all efforts to improve on the Lasso fit they are natural competitors to VISA. Our simulations contained five parameters that we altered. Namely, the number of variables (50 or 100), the number of observations (50 or 100), the number of non-zero coefficients (5 or 10), the values of the non-zero coefficients ($0.5, 0.75$ or 1) and the correlations among the columns in the design matrix ($0, 0.25, 0.4$ or 0.5). In the zero correlation case, the design matrices were generated using iid random Gaussian entries and we tested all combinations of the other parameters resulting in 24 simulations. For the correlated case, we fixed the non-zero coefficients at 1 and tested out all other combinations for an additional 24 simulations. In all 48 simulations, iid errors with a standard Normal distribution were added to the response variable. For each method and simulation we computed four statistics: False Positive, the number of variables with zero coefficients incorrectly included in the final model; False Negative, the number of variables with non-zero coefficients left out of the model; L2 square, the squared L_2 distance between the estimated coefficients and the truth; and MSE, the average prediction error of each method on a large test data set. We have not reported the MSE statistic, because it was generally very similar to the L_2 statistic. Tables 1 and 2 provide the other three statistics averaged over 200 data sets.

Table 1 reports results from a representative sample of six simulations with zero correlation and three with 0.5 correlation. All but the 50 variable, 100 observations with 0.5 correlation simulation used ten non-zero coefficients. These results illustrate the best possible performance of the various approaches with the tuning parameters chosen using the optimal point on the path, in terms of minimizing L2, for each of the seven methods. The tuning parameters were chosen individually for each of the 200 data sets. For the L2 statis-

Simulation	Statistic	VISA$_L$	VISA$_D$	DD	Relaxo	Adaptive	Dantzig	Lasso
50 var	False-Pos	1.84	2.01	2.35	2.24	6.27	13.78	15.32
100 obs	False-Neg	0.050	0.050	0.045	0.045	0.040	0.005	0.005
Coef= 0.5	L2-sq	**0.202**	**0.203**	0.218	0.216	0.290	0.373	0.356
50 var	False-Pos	0.23	0.43	0.46	0.39	2.98	13.71	15.84
100 obs	False-Neg	0	0	0	0	0	0	0
Coef= 1.0	L2-sq	**0.121**	**0.120**	0.129	0.128	0.173	0.366	0.345
100 var	False-Pos	2.97	3.23	4.16	3.88	26.48	14.97	18.72
100 obs	False-Neg	0.140	0.140	0.150	0.135	1.805	0.08	0.035
Coef= 0.5	L2-sq	**0.276**	**0.274**	0.305	0.302	1.174	0.553	0.516
100 var	False-Pos	0.49	0.61	0.90	0.72	33.80	15.10	18.66
100 obs	False-Neg	0	0	0	0	0.44	0	0
Coef= 1.0	L2-sq	**0.136**	**0.137**	0.154	0.149	1.84	0.592	0.539
100 var	False-Pos	8.07	9.96	9.73	8.45	*NA*	13.86	15.15
50 obs	False-Neg	2.380	1.805	2.180	2.340	*NA*	2.145	1.980
Coef= 0.5	L2-sq	1.256	**1.145**	1.211	1.270	*NA*	1.417	1.397
100 var	False-Pos	7.30	10.40	9.77	8.38	*NA*	15.00	17.52
50 obs	False-Neg	0.240	0.150	0.400	0.235	*NA*	0.425	0.185
Coef= 1.0	L2-sq	**1.264**	**1.211**	1.596	1.400	*NA*	2.891	2.254
50 var, 100 obs	False-Pos	0.49	0.82	1.14	0.57	1.46	9.74	8.57
Cor= .5	False-Neg	0	0	0	0	0.005	0	0
Coef= 1.0	L2-sq	**0.132**	**0.130**	**0.128**	0.137	0.160	0.306	0.314
100 var, 50 obs	False-Pos	11.30	13.61	11.46	10.84	*NA*	15.41	13.57
Cor= .5	False-Neg	0.190	0.130	0.425	0.340	*NA*	0.520	0.305
Coef= 1.0	L2-sq	1.852	**1.640**	1.984	2.032	*NA*	3.618	2.904
100 var, 100 obs	False-Pos	5.94	8.32	7.63	6.01	31.32	14.58	13.93
Cor= .5	False-Neg	0	0	0.020	0.005	0.815	0.055	0.005
Coef= 1.0	L2-sq	**0.439**	**0.454**	**0.450**	**0.448**	2.903	1.273	0.858

Table 1: *Simulation results using the optimal point on the path for each method.*

tic we performed tests of statistical significance, comparing each method to the best VISA approach. Since most differences were statistically significant, any differences between VISA and the other methods that were not significant were placed in bold font. For example, in the first simulation with 50 variables and 100 observations VISA$_L$ and VISA$_D$ were statistically identical, but the other methods were all significantly worse. In general, VISA$_L$ and VISA$_D$ performed similarly, with VISA$_L$ having slightly lower false positive rates but higher false negative rates. In comparison to the other methods, VISA generally had lower false positive rates and similar false negative. In terms of the L_2 error, the VISA methods were overall superior to all the other approaches in the zero correlation case, and in the correlated case were overall superior to everything except the Double Dantzig. In cases with more observations than variables, the Adaptive Lasso provided improved performance relative to the Dantzig selector and Lasso. However, because of its reliance on the least squares estimators, it performed poorly in other settings. The Double Dantzig and Relaxed Lasso performed somewhat similarly, and both provided considerable improvements over all methods except for VISA.

Simulation	Statistic	VISA$_L$	VISA$_D$	DD	Relaxo	Adaptive	Dantzig	Lasso
50 var	False-Pos	2.92	3.27	3.41	3.39	6.66	13.52	15.08
100 obs	False-Neg	0.045	0.045	0.070	0.065	0.065	0.005	0.005
Coef= 0.5	L2-sq	**0.226**	**0.228**	0.246	0.245	0.304	0.382	0.365
50 var	False-Pos	0.93	1.65	1.28	1.26	3.55	13.70	15.45
100 obs	False-Neg	0	0	0	0	0	0	0
Coef= 1.0	L2-sq	**0.141**	0.146	0.148	0.149	0.184	0.376	0.353
100 var	False-Pos	4.06	4.78	4.87	4.60	27.29	14.67	18.03
100 obs	False-Neg	0.155	0.145	0.170	0.170	1.805	0.075	0.040
Coef= 0.5	L2-sq	**0.303**	0.314	0.334	0.331	1.212	0.560	0.521
100 var	False-Pos	0.90	1.24	1.52	1.18	33.89	14.99	18.09
100 obs	False-Neg	0	0	0	0	0.465	0	0
Coef= 1.0	L2-sq	**0.153**	0.158	0.174	0.167	1.862	0.598	0.543
100 var	False-Pos	8.09	12.32	10.28	8.48	*NA*	13.40	15.05
50 obs	False-Neg	2.805	2.025	2.375	2.770	*NA*	2.365	2.135
Coef= 0.5	L2-sq	1.393	**1.333**	**1.352**	1.404	*NA*	1.461	1.438
100 var	False-Pos	7.72	11.44	10.53	9.10	*NA*	15.15	17.09
50 obs	False-Neg	0.265	0.200	0.390	0.245	*NA*	0.425	0.200
Coef= 1.0	L2-sq	**1.339**	**1.378**	1.672	1.469	*NA*	2.922	2.273
50 var, 100 obs	False-Pos	1.55	2.06	1.66	1.80	2.65	11.15	10.21
Cor= .5	False-Neg	0	0	0	0	0.005	0	0
Coef= 1.0	L2-sq	**0.161**	0.192	0.177	0.170	0.182	0.344	0.333
100 var, 50 obs	False-Pos	12.02	14.05	11.73	11.36	*NA*	15.37	13.70
Cor= .5	False-Neg	0.195	0.205	0.475	0.355	*NA*	0.515	0.300
Coef= 1.0	L2-sq	**1.985**	**1.982**	2.175	2.153	*NA*	3.708	2.919
100 var, 100 obs	False-Pos	7.18	10.05	8.47	7.06	31.67	14.57	14.46
Cor= .5	False-Neg	0	0	0.020	0.005	0.810	0.025	0.005
Coef= 1.0	L2-sq	**0.478**	0.601	**0.520**	**0.482**	2.953	1.333	0.866

Table 2: *Simulation results using a validation data set to choose the tuning parameters.*

In Table 2 we examine the deterioration in performance when the tuning parameter must also be chosen. For each of the 200 data sets in each simulation we produced a corresponding validation data set. The validation data sets were identically distributed to the training data and had the same number of observations and variables. We then selected the tuning parameters that gave the lowest mean squared error between the response and predictions on the validation data. As one would expect, this caused some deterioration in performance for all seven methods, but, with a few exceptions, the conclusions from Table 1 remain the same. Among the 24 zero correlation cases, VISA still generally outperformed the other methods, with VISA$_L$ being the best overall, however the advantage over the Double Dantzig and Relaxed Lasso was not as dramatic as for the optimal point. In the 24 correlated cases VISA$_L$ again produced the best results followed by the Double Dantzig.

We also performed three additional simulations with $p = 300$ predictors, $n = 50$ observations, ten non-zero coefficients and varying degrees of correlation and values of the coefficients. With these simulations we found VISA$_D$ was the best performer at low correlations. At higher correlation VISA$_L$ performed as well, followed by the Double Dantzig.

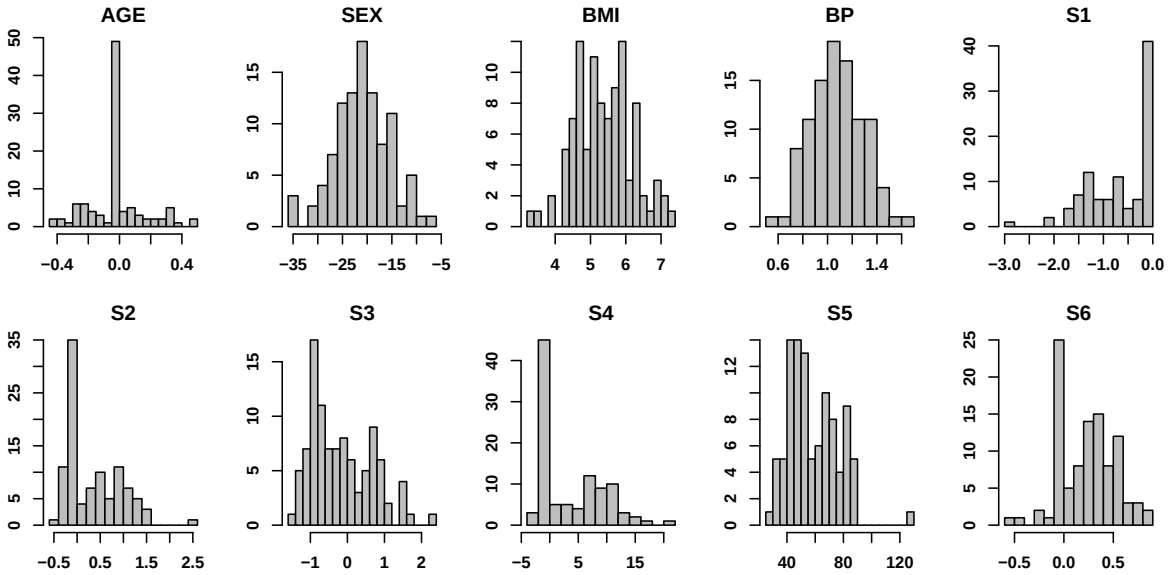

Figure 3: *Histograms of the* 100 *bootstrap coefficient estimates for the Diabetes data.*

4 Empirical Results

Here we present results from applying VISA to two real world data sets. We also provide VISA fits to a simple sparse simulated data set to evaluate performance in a setting where the truth is known. For all three data sets we implement VISA$_L$ using cross-validation to choose the optimal tuning parameters, λ and s. For a fixed value of λ, the VISA path for all values of s can be computed significantly faster than the whole LARS path. Of course, VISA uses two tuning parameters in comparison to LARS, which has only one. Hence, overall VISA is not quite as computationally efficient as LARS but the difference is generally fairly small. As a result, using cross-validation is quite feasible.

The first real world data set we examine is the Diabetes data used in the LARS paper of Efron *et al.* (2004). The data contains ten baseline predictors, age, sex, body mass, blood pressure and six blood serum measurements (S1,..., S6), for $n = 442$ diabetes patients. The response is a measure of disease progression one year after baseline. The average absolute pairwise correlation among the ten predictors is 0.31. As a result of the efficiency in computing the cross-validated VISA solution, we are able to implement a bootstrap approach, where the data is resampled, and the coefficients are estimated based on the resampled data. We use $B = 100$ bootstrap resamples and estimate the tuning parameters separately for each bootstrap data set. Histograms of the bootstrap estimates of the ten coefficients are presented in Figure 3. There appears to be strong evidence that Sex, BMI, BP and S5 are all statistically significant predictors of disease progression. In addition, there are clear spikes at zero for Age, S1, S2, S4 and to a slightly lesser extent S6. As opposed to standard linear regression, where one can only say that there is no evidence to include a variable, using VISA we see that there is actually evidence that these variables should be excluded. S3 is an interesting case. It has no spike at zero, indicating that it is being included in most

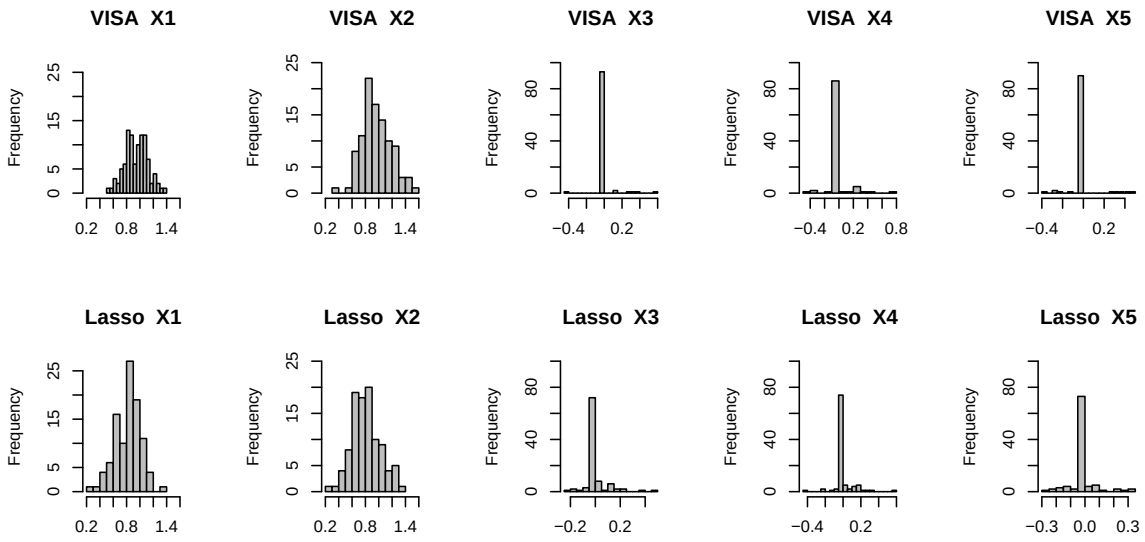

Figure 4: *Histograms of the coefficient estimates for the first five variables on* 100 *simulated data sets. VISA estimates are in the first row and Lasso in the second.*

of the bootstrap models but its coefficients fall either side of zero. This suggests that the variable may well be significant but it is unclear what effect it has on disease progression. One can also use a similar approach to compute the Lasso coefficient estimates. These fits show a similar pattern to those of VISA but with more spread in the insignificant variables and a lower spike at zero, resulting in a smaller signal.

Next we examine a simple simulated data set with ten predictors, generated from an iid standard Gaussian distribution, one hundred observations and normal errors with a standard deviation of two. The coefficients for X_1 and X_2 are both set to one while the remaining coefficients are zero. We first apply both VISA and Lasso to the data with the tuning parameters chosen using cross-validation. The resulting coefficient estimates are provided in Table 3. VISA has chosen the correct two variables with a small level of shrinkage of the coefficients. Lasso has had to choose between including too many variables or over shrinking the coefficients. In the end, the fit is a compromise with both, two additional variables included, and β_1 and β_2 over shrunk towards zero. The VISA coefficient vector is considerably more accurate, with a mean squared error relative to the true β equal to 0.006, compared to 0.032 for the Lasso. Additionally, when compared to the true response surface VISA has a mean squared error of 0.07 compared to 0.32 for Lasso. Next we generated 100 data sets with an identical distribution to the original, and applied VISA and the Lasso to each. Figure 4 plots histograms of the first five predictor coefficients for the two methods. The VISA coefficients for X_1 and X_2 are centered close to the truth of one, with means of 0.95 and 0.97, respectively. However, the Lasso coefficients are biased towards zero with means of 0.82 and 0.81, respectively. This bias also results in a considerably larger mean squared error for Lasso relative to VISA. In addition VISA produces many more zero estimates for X_3 through X_{10} with 86% estimated as zero, compared to only 69% for the

Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$
VISA	0.975	0.758	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Lasso	0.767	0.499	0.000	0.000	-0.126	0.000	0.000	0.000	0.000	0.056

Table 3: *VISA and Lasso coefficient estimates for the ten variable simulated data set.*

Order	Variable	Coefficient	Confidence Interval
1	Tuition	0.13	$(0.10, 0.16)$
2	Graduation Rate	5.61	$(1.69, 11.27)$
3	Expenditure per student	0.02	$(0.00, 0.04)$
4	Number of applications	0.06	$(0.02, 0.13)$
5	Cost of Books	0.48	$(0.08, 0.85)$
6	Parttime undergraduates	0.06	$(0.00, 0.17)$
7	Alumni donation rate	-11.72	$(-18.26, -4.47)$
8	Percent faculty with PhD	0.00	$(-5.70, 6.57)$

Table 4: *Coefficients and confidence intervals for the first eight variables to enter the model when fitting the college data. The variables are listed in the order that they entered.*

Lasso.

The final data set we examine is a subset of the USNews data used for the ASA 1995 Data Analysis Exposition. The data contains measurements on 18 variables from 777 colleges around the United States. The response of interest is the cost of room and board at each institution. The average absolute pairwise correlation among the 17 predictors is 0.32. We first randomly divide the data into a test data set of 100 observations, with the remainder making up the training data. Table 4 lists the first eight variables to enter the model, along with their coefficients and 95% bootstrap confidence intervals. The order that the variables enter provides a measure of their importance. We fit both VISA and Lasso to the training data and make predictions on the test data. The VISA predictions are statistically superior with a p-value of 0.02. Finally, Figure 5 provides plots of the cross-validated VISA errors for different values of λ and s. The dotted lines in the left hand plot demarcate decreasing values of λ, and hence larger models. Within each dotted region we have plotted the error for different values of s. Notice that for many values of λ, the optimal value of s is somewhere in the middle. The right hand plot provides the CV error, as a function of s, for the optimal λ. Again we see that the optimal s is not at the most extreme value. This provides further motivation for the VISA path approach as opposed to a simpler method where variable selection is performed first and then an OLS fit is used on the selected variables. On this data set, the entire set of cross-validated errors for all values of both tuning parameters took approximately four seconds to run in R.

We have opted here to compare VISA with Lasso because the Lasso is the most widely used of the methods we examine. From our simulation study we know that, while the improvement of VISA over the Double Dantzig or Relaxed Lasso is highly statistically significant, it is less extreme than that relative to the Lasso. Hence, had we compared VISA with the Double Dantzig or Relaxed Lasso one would not expect to have seen such dramatic differences.

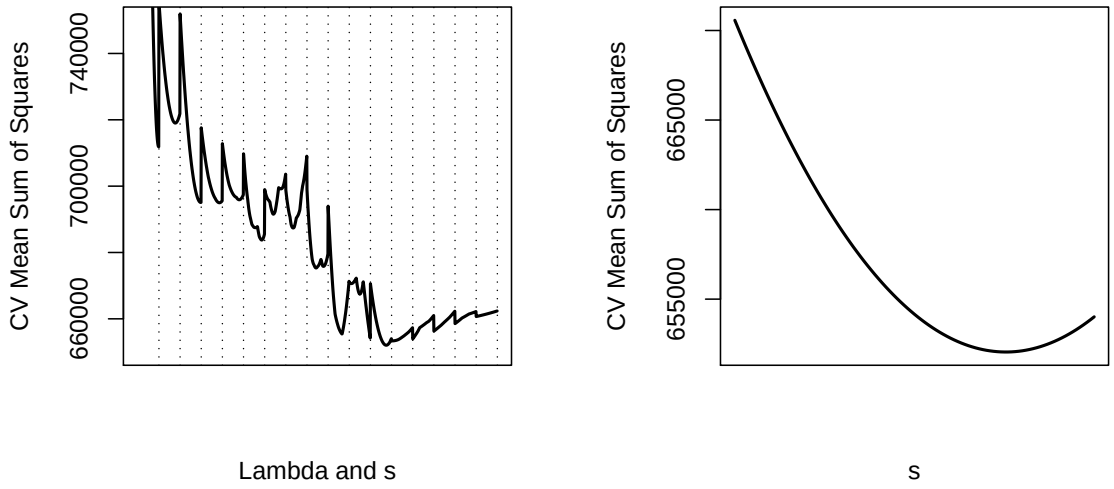

Figure 5: *Cross validated VISA error rates on the USNews data set for different values of* λ *and s.*

5 GLM VISA

Generalized linear models provide a framework for relating response and predictor variables (McCullagh and Nelder, 1989). For a random variable Y, we model the relationship between predictor $\mathbf{X}_i$ and response Y_i as $g(\mu_i) = \mathbf{X}_i^T \beta$, where $\mu_i = E(Y|\mathbf{X}_i)$ and g is referred to as the link function. Common examples of g include the identity link used for normal response data and the logistic link used for binary response data. In a standard GLM setup, the coefficient vector is generally estimated using maximum likelihood. However, when p is large relative to n, the maximum likelihood approach becomes undesirable for several reasons. First, maximum likelihood will not produce any coefficients that are exactly zero, and, as a result, the final model is less interpretable and probably less accurate. Second, for large p the variance of the estimated coefficients will become large and when $p > n$ there is no unique solution. To address these limitations several extensions of the Lasso to GLM data have been proposed (Park and Hastie, 2007). However, these extensions still suffer from over shrinkage of the coefficients. In this section we address this problem by extending VISA to the GLM setting.

5.1 GLM VISA Methodology

Notice that, for Gaussian error terms, $\mathbf{x}_j^T(\mathbf{Y} - X\beta) = \sigma l_j'(\beta)$ where l_j' is the partial derivative of the log likelihood function with respect to β_j, $\sigma^2 = var(\varepsilon_i)$ and $\mathbf{x}_j$ denotes the jth column of X. For a GLM model, using the canonical link, $l_j'(\beta) = \mathbf{x}_j^T(\mathbf{Y} - \mu)$ with $\mu = g^{-1}(X\beta)$. Hence, the VISA$_D$ optimization criteria, given by (5) and (6), can be extended in a natural

fashion by computing the $\beta_\lambda(s)$ that minimizes $\|\tilde{\beta}\|_1$ subject to

$$\begin{aligned} \|X^T(Y-\tilde{\mu})\|_\infty &\leq \lambda, \quad \text{for } s \text{ in } [0,2\lambda] \\ \|X^T_{\mathcal{A}_1}(Y-\tilde{\mu})\|_\infty &\leq \lambda - s, \quad \text{for } s \text{ in } (0,\lambda] \qquad (9) \\ \|X^T_{\mathcal{A}_1}(Y-\tilde{\mu})\|_\infty = 0, \quad \|X^T_{\mathcal{A}_2}(Y-\tilde{\mu})\|_\infty &\leq 2\lambda - s, \text{ for } s \text{ in } (\lambda,2\lambda]. \qquad (10) \end{aligned}$$

Here we define $\mathcal{A}_1$ and $\mathcal{A}_2$ in an analogous fashion to the definitions for the linear case, except that $X\beta$ is replaced by μ. For the Gaussian distribution with the identity link function, the GLM VISA reduces to the standard VISA but it can also be applied to a much wider range of response distributions. For example, solving (9) and (10) using a logistic link function allows us to perform regression for categorical $\{0,1\}$ response data.

Unless g is the identity function, the GLM VISA constraints are not linear, so linear programming software can not be directly used to compute $\beta_\lambda(s)$. However, recall that in the standard GLM setting, an iterative weighted least squares algorithm is used to solve the likelihood equation. In particular, given a current estimate $\tilde{\beta}$, an adjusted dependent variable $Z_i = \mathbf{X}_i^T\tilde{\beta} + (Y_i - \tilde{\mu}_i)/V_i$ is computed, where V_i is the conditional variance of Y_i given $\mathbf{X}_i$. A new estimate for β is then produced using weighted least squares, i.e. by solving $\mathbf{x}_j^T W(\mathbf{Z} - X\tilde{\beta}) = 0$ for $j = 1,\ldots,p$, where W is a matrix consisting of the V_i's on the diagonal. This procedure is iterated until $\tilde{\beta}$ converges.

An analogous iterative approach can be used to compute the GLM VISA.

1. At the $k+1$th iteration, let V_i equal the conditional variance of Y_i given the current parameter estimates, and define $Z_i = \mathbf{X}_i^T\beta^{(k)} + (Y_i - \mu_i^{(k)})/V_i$, where (k) denotes the corresponding estimate from the kth iteration.

2. Let $Z_i^* = Z_i\sqrt{V_i}$ and $X_{ij}^* = X_{ij}\sqrt{V_i}$.

3. Optimize (5) and (6) to compute $\beta^{(k+1)}$ using $\mathbf{Z}^*$ as the response and X^* as the design matrix.

4. Repeat steps 1 through 3 until convergence.

As mentioned previously, (5) and (6) can be formulated as linear programming problems, thus step 3 can be computed efficiently. James and Radchenko (2008) use a similar algorithm and note that it generally converges within a few iterations, in which case the computation time for the GLM VISA would only be a small multiple of that for the standard VISA.

In addition to directly computing the GLM VISA solution for each given λ and s, we can also adapt the VISA algorithm to generate the entire path of GLM VISA coefficients. Our path algorithm is a natural generalization of the algorithms given in Park and Hastie (2007) and James and Radchenko (2008). Although GLM VISA coefficient paths are not piece-wise linear, they can be well approximated by a piece-wise linear solution. Given a point on the path, we use the linear approximation to the GLM VISA optimization problem together with the steps 2 and 3 of the VISA$_D$ algorithm to approximately identify the

next point on the path where a coefficient becomes nonzero, a coefficient hits zero, or a new constraint becomes active. Then we use the new coefficient vector to iteratively solve the corresponding GLM VISA optimization problem, as described above. We call this a correction step. The fact that we know the set of active constraints and the set of nonzero coefficients allows us to avoid using an optimization package in the correction step, and instead solve a system of linear equations at each iteration.

5.2 Non-Asymptotic Bound

Similar non-asymptotic bounds to those for standard VISA, exist for GLM VISA. In particular, let $\mathcal{H}$ be the parameter space of coefficients, β. Further, let $\mathcal{D}$ be the set of all diagonal matrices with ith diagonal entry $(g^{-1}(\mathbf{X}_i^T\alpha) - g^{-1}(\mathbf{X}_i\beta))/\mathbf{X}_i^T[\alpha-\beta]$ where $\alpha \in \mathcal{H}$.

Definition 2 *Let $\psi(k)$ denote the smallest eigenvalue of the matrices in $\{X_J^T D X_J, |J| \leq k, D \in \mathcal{D}\}$.*

Then Theorem 2 extends the non-asymptotic bounds from Section 2.4 to the GLM domain.

Theorem 2 *Suppose that $\beta \in \mathbb{R}^p$ is an S-sparse coefficient vector. Let $\hat{\beta}$ be a GLM VISA estimator with k false positive coefficients, and let $\lambda_\infty = \|X^T(Y-\hat{\mu})\|_\infty$. Then,*

$$P\left(||\hat{\beta}-\beta||_2 > \frac{(S+k)^{1/2}}{\psi(S+k)}(\lambda_\infty+\tau)\right) \leq P(\xi > \tau)$$

for each positive τ, where $\xi = ||X^T(Y-\mu)||_\infty$.

When $\psi(S+k)=0$, we assume that the probability on the left-hand side of the above inequality equals zero. In the Gaussian case, $\psi \equiv \phi$, and $P(\xi > a\sqrt{\log p}) \leq (p^a\sqrt{4\pi\log p})^{-1}$, as shown in the Appendix. In the more general setting Theorem 3 places a bound on $P(\xi > \tau)$ for most typical response distributions.

Theorem 3 *Fix a positive a.*

1. *Suppose that there exist positive constants K and v, such that the response variables satisfy $K\left(Ee^{|Y_i-\mu_i|^2/K}-1\right) \leq v$ for every i. Set $\tau_p = a\sqrt{\log p}$ and define $\gamma = a^2[8K+8v]^{-1}$. Then*
$$P(\xi > \tau_p) \leq 2p^{1-\gamma}.$$

2. *Suppose that there exist positive constants M and v, such that the response variables satisfy $2M^2E\left(e^{|Y_i-\mu_i|/M}-1-\frac{|Y_i-\mu_i|}{M}\right) \leq v$ for every i. Set $\tau'_p = a\log p$ and define $\gamma = a[2M+2v/(a\log p)]^{-1}$. Then*
$$P(\xi > \tau'_p) \leq 2p^{1-\gamma}.$$

The first part of Theorem 3 covers, for example, the normal distribution with bounded variance and the binomial distribution. The second part covers, for example, the poisson and the exponential distributions, both with the assumption of bounded variance. Combining Theorems 2 and 3 allows us to construct bounds on the L_2 error in the coefficient estimates, which hold with high probability, for all the common response distributions. For example, consider the binomial distribution with $Y_i = 0$ or 1. Then the condition in part 1 holds, for example, with $K = 1/2$ and $v = 1/2$. Consequently, inequality

$$||\hat{\beta} - \beta||_2 \leq \frac{(S+k)^{1/2}}{\psi(S+k)}(\lambda_\infty + 4\gamma^{1/2}\sqrt{\log p})$$

is satisfied for each positive γ with probability at least $1 - 2p^{1-\gamma}$.

6 Discussion

The Lasso, Dantzig selector, Relaxed Lasso and Double Dantzig, all use a hard thresholding rule to select the model variables. The Relaxed Lasso and Double Dantzig then use a second tuning parameter to adjust the level of shrinkage on the selected variables. Our simulation results, along with previous studies, demonstrate that this two stage approach can produce considerable improvements over the Lasso and Dantzig selector. However, the hard thresholding rule means that none of these approaches can correct any mistakes in the initial model selection step. The key contribution of the VISA methodology is to introduce a more flexible selection scheme, where variables can potentially enter or leave the model as the fit to the data improves. Our simulation results show that this more flexible strategy can produce considerable improvements over the Lasso and Dantzig selector as well as smaller, but still statistically significant, improvements over the Relaxed Lasso and Double Dantzig. In addition to its strong practical performance VISA also possesses interesting theoretical properties, that suggest it should perform well as p tends to infinity. Finally, the standard VISA methodology can be extended to the class of GLM response distributions, providing an added level of flexibility.

We see a few possible future directions for this work. VISA is essentially performing a three step procedure, where first a primary set of variables is identified, then a secondary set, and finally the optimal level of shrinkage is chosen. However, this idea could be extended beyond three steps. For example, one could use the same path approach to implement a four step procedure by selecting a tertiary set of variables. In theory we could keep adding steps, but one may imagine that there are diminishing levels of return. In addition, we currently use the same cutoff for determining both the primary and secondary variables. The advantage of this approach is that it can be implemented with only two tuning parameters. However, the results in Lemma 1 could be strengthened with the addition of a third tuning parameter to allow a different cutoff for the secondary variables. It is an open question whether any potential improvement in the practical performance of VISA would outweigh the difficulty of selecting a third tuning parameter.

Acknowledgements

We would like to thank the Associate Editor and referee for many helpful suggestions that improved the paper. This work was partially supported by NSF Grant DMS-0705312.

A Steps 2 and 3 of the VISA$_D$ algorithm

Write β^+ and β^- for the positive and negative parts of β^l. Suppose that the indexes in $\mathcal{A}$ (and, correspondingly, in $\mathcal{A}_\downarrow^M$) are ordered according to the time they were added to $\mathcal{A}$. Let $\tilde{I}$ be an $|\mathcal{A}|$ by $|\mathcal{A}_\downarrow^M|$ matrix in which the ij'th element equals one if the i'th member of $\mathcal{A}$ is also the j-th member of $\mathcal{A}_\downarrow^M$, and it equals zero otherwise. Write S for the $|\mathcal{A}|$-dimensional diagonal matrix containing the signs of covariances c_j for the variables in $\mathcal{A}$, and compute the $|\mathcal{A}|$ by $2p+|\mathcal{A}_\downarrow^M|$ matrix $A = \left[\begin{array}{ccc} -SX_{\mathcal{A}}^T X & SX_{\mathcal{A}}^T X & \tilde{I} \end{array}\right]$. The first p columns of A correspond to β^+ and the next p columns to β^-. Let $\tilde{B}$ be the matrix produced by selecting all the columns of A that correspond to the non-zero components of β^+ and β^-, and let A_i be one of the remaining columns. Write the two matrices in the following block form:

$$\tilde{B} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \qquad \text{and} \qquad A_i = \begin{pmatrix} A_{i_1} \\ A_{i_2} \end{pmatrix},$$

where B_1 is a square matrix of dimension $|\mathcal{A}|-1$, and A_{i_2} is a scalar. Define

$$i^* = \underset{i:\,|q_i|\neq 0,\,\alpha/q_i>0}{\arg\max} \left[\mathbf{1}^T B_1^{-1} A_{i_1} - 1_{\{i\leq 2p\}}\right] / |q_i|, \tag{11}$$

where $q_i = A_{i_2} - B_2 B_1^{-1} A_{i_1}$, $\alpha = B_2 B_1^{-1}\tilde{\mathbf{1}} - \tilde{1}$, and $(\tilde{\mathbf{1}}^T, \tilde{1})$ is a zero-one row vector of length $|\mathcal{A}|$ that indicates whether the corresponding element of $\mathcal{A}$ belongs to $\mathcal{A}_\downarrow^M$.

If $i^* \leq 2p$, augment the set $\mathcal{B}$ by the index of the corresponding variable. If $i^* = 2p+m$ for $m = 1, \ldots, |\mathcal{A}_\downarrow^M|$, leave $\mathcal{B}$ unchanged, but remove the m-th element from the set $\mathcal{A}_\downarrow^M$.

As with VISA$_L$, the first point at which a new $|c_j|$ reaches λ is given by

$$\gamma_1 = \overset{+}{\min}_{j\in\mathcal{A}_=^c\setminus\mathcal{A}_\downarrow} \left\{ \frac{c_j-\lambda}{\mathbf{x}_j^T X\mathbf{h}}, \frac{c_j+\lambda}{\mathbf{x}_j^T X\mathbf{h}} \right\}.$$

Let $\mathbf{x}_k$ be a variable that is a member of the set $\mathcal{A}_\downarrow^M$. Then, as with LARS, the first point a new $|c_m|$ hits C for m in $\mathcal{A}_\downarrow$ is given by

$$\gamma_2 = \overset{+}{\min}_{j\in\mathcal{A}_\downarrow\setminus\mathcal{A}_\downarrow^M} \left\{ \frac{c_k-c_j}{(\mathbf{x}_k-\mathbf{x}_j)^T X\mathbf{h}}, \frac{c_k+c_j}{(\mathbf{x}_k+\mathbf{x}_j)^T X\mathbf{h}} \right\},$$

and the first point that non-zero coefficient crosses zero is given by $\gamma_3 = \min_j^+ \left\{-\beta_j^l/h_j\right\}$. Combining the above with the possibility that $C=0$, we get $\gamma = \min\{\gamma_1, \gamma_2, \gamma_3, c_k/\mathbf{x}_k^T X\mathbf{h}\}$.

B Proof of Lemma 1

Write Z_j for $\mathbf{x}_j^T\varepsilon$ and note that $Z_i \sim N(0,\sigma^2)$ for each i. Absolute covariances at the start of the path are given by

$$\begin{aligned}
|c_1| &= |\mathbf{x}_1^T\mathbf{Y}| = |\beta_1 + \rho_{12}\beta_2 + Z_1| \\
|c_2| &= |\mathbf{x}_2^T\mathbf{Y}| = |\rho_{12}\beta_1 + \beta_2 + Z_2| \\
|c_j| &= |\mathbf{x}_j^T\mathbf{Y}| = |\rho_{1j}\beta_1 + \rho_{2j}\beta_2 + Z_j| \quad \text{for } j \in J_n.
\end{aligned}$$

Observe that $|c_1| = \beta_1 + o_p(n^{1/2})$ and $|c_2| < (1-\delta)\beta_1 + o_p(n^{1/2})$. Note the stochastic bound $\max_{J_n}|Z_j| = O_p(\sqrt{\log p})$, and deduce that $\max_{J_n}|c_j| < (1-\delta)\beta_1 + o_p(n^{1/2})$. Thus, with probability tending to one, the first variable selected by LARS is X_1.

The first part of the LARS algorithm drives the absolute covariance $|c_1|$ towards zero using one non-zero coefficient: $\hat{\beta}_1$. Absolute covariances on this part of the LARS path are given by

$$\begin{aligned}
|c_1| &= |\mathbf{x}_1^T(\mathbf{Y} - \hat{\beta}_1\mathbf{x}_1)| = |(\beta_1 - \hat{\beta}_1) + \rho_{12}\beta_2 + Z_1| \\
|c_2| &= |\mathbf{x}_2^T(\mathbf{Y} - \hat{\beta}_1\mathbf{x}_1)| = |\rho_{12}(\beta_1 - \hat{\beta}_1) + \beta_2 + Z_2| \\
|c_j| &= |\mathbf{x}_j^T(\mathbf{Y} - \hat{\beta}_1\mathbf{x}_1)| = |\rho_{1j}(\beta_1 - \hat{\beta}_1) + \rho_{2j}\beta_2 + Z_j| \quad \text{for } j \in J_n.
\end{aligned}$$

If $|c_2|$ is the first absolute covariance that rises up to $|c_1|$, then simple algebra shows that when $|c_2|$ becomes active,

$$|c_2| = (1+\rho_{12})\beta_2 + o_p(\beta_2). \tag{12}$$

If $|c_j|$ with $j > 2$ is the first absolute covariance that rises up to $|c_1|$, then when $|c_j|$ becomes active,

$$|c_j| = \Big(\frac{\rho_{2j} - \rho_{1j}\rho_{12}}{1-\rho_{1j}}\Big)\beta_2 + o_p(\beta_2). \tag{13}$$

Conclude that under the assumptions of part 1 of the lemma, the second variable selected by LARS is X_2, with probability tending to one. Set λ to the maximum absolute covariance at this point, $\lambda = (1+\rho_{12})\beta_2 + o_p(\beta_2)$, and consider the corresponding VISA path $\hat{\beta}_L(\lambda,\cdot)$. VISA starts by driving $|c_1|$ and $|c_2|$ towards zero at the same rate, using the coefficients $\hat{\beta}_1$ and $\hat{\beta}_1$. Absolute covariances $|c_j|$ for the noise variables in J_n are given by

$$|\mathbf{x}_j^T(\mathbf{Y} - \hat{\beta}_1\mathbf{x}_1 - \hat{\beta}_2\mathbf{x}_2)| = |\rho_{1j}(\beta_1 - \hat{\beta}_1) + \rho_{2j}(\beta_2 - \hat{\beta}_2) + Z_j| = \frac{\rho_{1j} + \rho_{2j}}{1+\rho_{12}}|c_1| + O_p(1),$$

hence $\max_{J_n}|c_j| < (1-\delta_1/2)\lambda + o_p(\beta_2)$. Conclude that with probability tending to one, none of the $|c_j|$'s will rise up to the level λ throughout the VISA path. This completes the proof of part 1, because the final point of the coefficient path is the solution to the system of equations $\mathbf{x}_1^T(\mathbf{Y} - \hat{\beta}_1\mathbf{x}_1 - \hat{\beta}_2\mathbf{x}_2) = 0$ and $\mathbf{x}_2^T(\mathbf{Y} - \hat{\beta}_1\mathbf{x}_1 - \hat{\beta}_2\mathbf{x}_2) = 0$, namely β_{12}^{ols}.

Recall the assumptions of part 2 of the lemma and compare the expressions for the absolute covariances in equations (12) and (13). Conclude that with probability tending to one, LARS will select a noise variable ahead of X_2. Set λ to the maximum absolute

covariance right before the noise variable enters the model:

$$\lambda = \max_{j \in J_n} \frac{\rho_{2j} - \rho_{1j}\rho_{12}}{1 - \rho_{1j}} \beta_2 + o(\beta_2).$$

Consider the corresponding VISA path $\hat{\beta}_L(\lambda, \cdot)$. The fact that the correlation between each noise variable and X_1 is positive guarantees that as $|c_1|$ is decreased, the maximum absolute covariance for the noise variables remains below the level λ with probability tending to one. If $|c_1|$ gets driven all the way down to zero, the corresponding value of $\beta_1 - \hat{\beta}_1$ is $-\rho_{12}\beta_2 + o_p(\beta_2)$, hence the corresponding value of $|c_2|$ is $(1-\rho_{12}^2)\beta_2 + o_p(\beta_2)$. Compare the latter value to λ taking into account inequality (7) and conclude that, with probability tending to one, the absolute covariance $|c_2|$ will reach the level λ at some point on the VISA path before $|c_1|$ hits zero. Thus, VISA selects X_2 as a secondary variable. Again, the fact that all the noise variables are positively correlated with X_1 and X_2 guarantees that $\max_{J_n} |c_j|$ will remain below the level λ throughout the VISA path. Complete the proof by noting that the final point of the path is again the oracle least squares solution β_{12}^{ols}.

C Proof of Theorem 1

Define $h = \hat{\beta} - \beta$, let J be the index set of the $S+k$ nonzero elements of h, and write h_J for the corresponding subvector of h. Note that

$$\|X_J^T X_J h_J\|_\infty = \|X_J^T X(\hat{\beta} - \beta)\|_\infty = \|X_J^T(\mathbf{Y} - X\beta) - X_J^T(\mathbf{Y} - X\hat{\beta})\|_\infty \le \lambda_\infty + \|X^T\varepsilon\|_\infty.$$

Hence the inequality $\|X^T\varepsilon\|_\infty \le \tau_p$ implies

$$\lambda_\infty + \tau_p \ge \|X_J^T X_J h_J\|_\infty \ge (S+k)^{-1/2}\|X_J^T X_J h_J\|_2 \ge (S+k)^{-1/2}\phi(S+k)\|h_J\|_2,$$

thus the bound $\phi(S+k)\|h\|_2 \le (S+k)^{1/2}(\lambda_\infty + \tau_p)$ holds with probability at least $P(\|X^T\varepsilon\|_\infty \le \tau_p)$. Note that each component of the vector $X^T\varepsilon$ has a $N(0,\sigma^2)$ distribution, hence the desired probability bound $P(\|X^T\varepsilon\|_\infty > \tau_p) \le (p^a\sqrt{4\pi\log p})^{-1}$ follows from standard results for normal random variables.

D Proof of Theorem 2

Note that vector $X^T(\hat{\mu} - \mu)$ can be written as $X^T D X(\hat{\beta} - \beta)$ for some matrix D in the set $\mathcal{D}$. Argue as in the proof of Theorem 1 to establish $\psi(S+k)\|\hat{\beta} - \beta\|_2 \le (S+k)^{1/2}(\lambda_\infty + \tau)$ with probability at least $P(\|X^T(Y-\mu)\|_\infty \le \tau)$.

E Proof of Theorem 3

We first present two lemmas and a corollary used in the proof. The following lemma can be found in Bennett (1962), pages 37-38.

Lemma 2 (Bernstein's inequality) *Let V_1,...,V_n be independent random variables with zero mean such that inequalities $E|V_i|^m \leq m!M^{m-2}v_i/2$ hold for every $m \geq 2$ (and all i) and some positive M and v_i. Then*

$$P\big(|V_1+...+V_n|>x\big) \leq 2\exp\left[-\frac{x^2}{2\big(Mx+\sum_{i=1}^n v_i\big)}\right].$$

The next result is a direct consequence.

Corollary 2 *Suppose that W_1,...,W_n are independent random variables with expectation zero and with $\max_{i\leq n} 2M^2E\big(e^{|W_i|/M}-1-\frac{|W_i|}{M}\big) \leq v$ for some positive M and v. Then for all positive x and real a_i with $|a_i| \leq 1$,*

$$P\big(|a_1W_1+...+a_nW_n|>x\big) \leq 2\exp\left[-\frac{x^2}{2\big(Mx+v\sum_{i=1}^n a_i^2\big)}\right].$$

Proof: For each i, write out the Taylor expansion for the left-hand side of inequality $2M^2E\big(e^{|W_i|/M}-1-\frac{|W_i|}{M}\big) \leq v$ to derive $E|W_i|^m \leq m!M^{m-2}v/2$ for every $m \geq 2$. Define $V_i = a_iW_i$ and note that $E|V_i|^m \leq m!M^{m-2}a_i^2v/2$. Apply Bernstein's inequality to complete the proof.

The following lemma is from Van de Geer (1999), Lemma 8.2.

Lemma 3 *Suppose that $W_1,\ldots,W_n$ are independent random variables with expectation zero and with $\max_{i\leq n} K(Ee^{|W_i|^2/K}-1) \leq v$ for some positive K and v. Then for all real a_i and $x>0$,*

$$P\big(|a_1W_1+...+a_nW_n|>x\big) \leq 2\exp\left[-\frac{x^2}{8(K+v)\sum_{i=1}^n a_i^2}\right].$$

Theorem 3 follows from the above results. Let $\mathbf{X}_i$ denote the i-th column of matrix X and write W for $Y-\mu$. Observe that $P\big(||X^TW||_\infty > x\big)$ is bounded above by $\sum_{i=1}^p P\big(|\mathbf{X}_i^TW|>x\big)$. Bound each summand by $2p^{-\gamma}$ by applying either Lemma 3 with $x=c\sqrt{\log p}$ or Corollary 2 with $x=c(\log p)$, depending on the behavior of the tails of the response distribution.

References

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **57**, 33–45.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Annals of Statistics* **35**, 6, 2313–2351.

Chen, S., Donoho, D., and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM Journal of Scientific Computing* **20**, 33–61.

Efron, B., Hastie, T., Johnston, I., and Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics* **32**, 2, 407–451.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 2, 109–135.

James, G. M. and Radchenko, P. (2008). A generalized dantzig selector with shrinkage tuning. *Under review. (Available at www-rcf.usc.edu/∼gareth)* .

James, G. M., Radchenko, P., and Lv, J. (2008). The DASSO algorithm for fitting the dantzig selector and the lasso. *Under review. (Available at www-rcf.usc.edu/∼gareth)* .

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2nd edn.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis* **52**, 374–393.

Park, M. and Hastie, T. (2007). An l1 regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B* **69**.

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

Van de Geer, S. (1999). *Empirical Processes in M-Estimation*. Cambridge University Press.

Zhao, P., Rocha, G., and Yu, B. (2006). Grouped and hierarchical model selection through composite absolute penalties. Technical Report, Department of Statistics, University of California at Berkeley.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

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