MAGIC: a general, powerful and tractable method for selective inference

Xiaoying Tian, Nan Bi and Jonathan Taylor

Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305, USA
e-mail: xytian@stanford.edu
nbi@stanford.edu
jonathan.taylor@stanford.edu

Abstract: Selective inference is a recent research topic that tries to perform valid inference after using the data to select a reasonable statistical model. We propose MAGIC, a new method for selective inference that is general, powerful and tractable. MAGIC is a method for selective inference after solving a convex optimization problem with smooth loss and $\ell_1$ penalty. Randomization is incorporated into the optimization problem to boost statistical power. Through reparametrization, MAGIC reduces the problem into a sampling problem with simple constraints. MAGIC applies to many $\ell_1$ penalized optimization problem including the Lasso, logistic Lasso and neighborhood selection in graphical models, all of which we consider in this paper.

Keywords and phrases: selective inference, statistical power, sampling.

1. Introduction

There are a great deal of sophisticated statistical learning methods that allow us to search through a large number of models and look for meaningful patterns. Having done this search, we naturally want to judge the apparent associations that have been found. This has spawned a new area of research called selective inference [5, 6, 14, 15]. Loosely speaking selective inference recognizes the inherent selection biases in reporting the most “significant” results from various statistical models and attempts to adjust for the bias.

At a high level, selective inference involves two stages: First, solve a convex optimization problem, usually some penalized loss function. Second, perform inference in the statistical model suggested by the result of the optimization problem. For example, we first use the data to solve the Lasso problem, and then want to form confidence intervals for the variables that are nonzero in the Lasso solution. Adjustment for selection results in some constraints on the underlying distribution. Although various such problems have been studied, most of the papers only focus on one specific optimization problem. This is necessary as different loss functions in the optimization problems result in different geometry of the constraints. In this paper, we introduce a method, called ”MAGIC”, Monte-carlo Algorithm for General Inference with Constraints, which provides valid selective inference for optimization problems with any

*Supported in part by NSF grant DMS 1208857 and AFOSR grant 113039.
smooth loss functions. The advantage of MAGIC compared to previous selective inference methods are generality, statistical power and tractability. We elaborate each in the following passage:

**Generality:** The generality of MAGIC lies in two aspects: arbitrary smooth loss function in the penalized optimization problem and the data distribution from any exponential family. In comparison, the authors in [6] considered only inference after solving Lasso; The work [5] considered some exponential families with simple selection rules, but also noticed the difficulty for inference after solving more complex optimization problems. Finally, the work [13] is the closest in generality to this work, but shows substantially weaker statistical power, which we discuss below.

**Statistical power:** Earlier work [6, 13] has provided valid inference after selection, but sometimes lacks power. Other work [15, 5] suggested introducing randomness in the optimization algorithm which produces much improved power. This is the approach we take in this work. In simulation, we show that MAGIC produces much improved power over [6, 13].

**Tractability:** The earlier work [6, 13] computes valid p-values in closed form, thus involving the least computation cost. The framework in [5, 15] involves sampling in a constrained subset in the sample space. Both work used hit-and-run algorithm proposed in [1], which is a method to generate distributions on a subset of the space. The constrained subsets as described in [5, 15, 6, 13] can be quite complicated and depend on the particular loss function. Algorithms that do not use MCMC, such as [6, 13], do not suffer from this problem too much as they only need to compute the boundary once, but the methods in [5, 15] need to compute the boundary at each step of simulation, resulting in much more computation cost. MAGIC, however, transforms the constrained subset to a canonical set through reparametrization, removing the computational cost involved in computing the boundary at each step of sampling. Thus it is more tractable than [5, 15].

In Section 2, we introduce the general form of our randomized optimization problem, and describe the inference method as well as theory for MAGIC. Section 3 gives applications of MAGIC to different statistical learning problems. To demonstrate the applicability of MAGIC, we give three distinct examples: Lasso, $\ell_1$ penalized logistic regression and neighbourhood selection [8], which are applied in regression, classification and Gaussian graphical models respectively. Section 4 includes the comparisons of MAGIC with existing selective inference methods both in terms of statistical power and confidence intervals. All proofs are collected in Section 5 and the sampling methods are covered in Section 6. We conclude with discussions about future work in Section 7.

### 1.1. Related works

Most of the theoretical work on high-dimensional data focuses on consistency, either the consistency of solutions [11, 17] or the consistency of the models [19, 21].

In the post selection literature, the authors in [2] proposed the PoSI approach, which reduce the problem to a simultaneous inference problem. Because of the simultaneity, it prevents data snooping from any selection procedure, but also results in more conservative inference. In addition, the PoSI method has extremely high computational
cost, and is only applicable when the dimension $p < 30$ or for very sparse models. The authors [10] proposed a method for computing p-values that controls false discovery rate (FDR) among all variables. This is quite different from the hypothesis testing framework of this work, as the hypotheses tested in selective inference are chosen as a function of the data. Hence, the hypotheses tested are not directly comparable. Furthermore, compared with [10], MAGIC has the advantage of being able to construct confidence intervals for the selected variables.

2. Randomized selective inference

2.1. A randomized selection algorithm

Many statistical learning problems can be cast as convex optimization problems. Specifically, we solve the following randomized convex optimization.

$$\hat{\beta} = \min_{\beta \in \mathbb{R}^p} \ell(\beta; S) + \lambda \|\beta\|_1 - \omega^T \beta,$$

where data $S \sim F$, $\ell$ can be the negative log-likelihood for $F$, but generally just needs to be some convex loss function in $\beta$, the randomization variable $\omega \sim G$, a distribution on $\mathbb{R}^p$ independent of $F$, $\lambda$ is fixed. Without randomization, that is to set $G = \delta_0$, the point mass at 0, (1) includes many classical statistical learning problems, e.g. lasso [16], elastic net [22], $\ell_1$ penalized logistic regression, neighbourhood selection [8]. Although it might seem strange to add noise to data for model selection, it is seen in other forms in literature and applications. Common use of data splitting is an example [4, 20], as a random subset of data is used for model selection. The form of our randomization is also related to [9]. We can control the amount of randomization through the variance of $G$, usually just a little randomization will produce much improved power.

We define the variable selection map as

$$E(S, \omega) = \text{supp}(\hat{\beta}(S, \omega)).$$

For the observed data $S_{\text{obs}}$ and an instance of $\omega_{\text{obs}}$, both considered fixed, we define $E = E(S_{\text{obs}}, \omega_{\text{obs}})$ which is the active set of (1) and consider it fixed hereafter.

After having solved the above problem, we now consider inference for parameters chosen on the basis of this set of non-zero coefficients $E$. Suppose the data $S \sim F$ is a member of an exponential family with parameters $b \in \mathbb{R}^p$ and sufficient statistics $T(S) \in \mathbb{R}^p$. In particular, its density $f_b(s)$ has the following form,

$$\frac{df_b}{d\mu}(s) = \exp(b^T T(s) - A(b))$$

where $\mu$ is the reference measure on the sample space of $S$ and $A$ is the normalizing constant with $\mu, A$ known. Having observed a set of selected variables $E$, we can and often do then consider a submodel of the above model with $b_{-E} = 0$. If $E \supseteq \text{supp}(b)$, then our model is correctly specified. This is the scenario we always consider hereafter. For treatment of misspecified models, see [6, 5]. Under this submodel, the joint distribution of $(S, \omega)$ is fully specified. Our target of inference is now $b_E$. 
Since $E$ is not given a priori, but selected by the data, it seems to be only fair to consider $(S, \omega)$ such that $\hat{E}(S, \omega) = E$. This is equivalent to condition on the event \{(S, \omega) : \hat{E}(S, \omega) = E\}. This is the general approach taken in [6, 5, 15] to provide valid (selective) inference in the above model.

Let $A$ be the region where \{(S, \omega) \in A \iff \hat{E}(S, \omega) = E\}, then this general approach to selective inference requires us to describe the conditional distribution $S | (S, \omega) \in A$, $(S, \omega) \sim F \times G$. (2)

We first state the following result,

**Theorem 1.** Suppose $(S, \omega) \sim F \times G$, $F$ is the exponential family specified above, with the parameters $b$ satisfying $\text{supp}(b) \subseteq E$. $G$ is a distribution on $\mathbb{R}^p$ and $A$ is defined as above. Then for any variable $j \in E$, there exists a p-value function $P_j : \text{supp}(F) \rightarrow [0, 1]$, such that $
abla F \times G[P_j(S; A) \leq \alpha | (S, \omega) \in A] \leq \alpha$, (3) under the null hypothesis $H_{0j} : b_j = 0$. The function $P_j$ only depends on data $S$ and $A$.

In some cases, equality holds in (3), we will discuss the conditions in the proof. In this case, the test proposed above is the Uniformly Most Powerful Unbiased test [5], providing theoretical ground for the power of MAGIC. Theorem 1 gives a construction of the p-value, which we can use to reject the null hypothesis at level $\alpha$. We will give the exact construction of $P_j$ in the proof of Theorem 1, which is an multivariate integral and is hard to compute in general. We instead try to acquire samples from (2) and approximate the multivariate integral. The constrained region $A$ is the bottleneck for the sampling, which is complicated and specific to the loss function $\ell$. However, through a reparametrization, we can form the constrained region as a simple set that is independent of $\ell$.

### 2.2. Augmented parameter space

Once we solve the optimization (1), we get $\hat{\beta}$ the solution and $\hat{z}$ the subgradient of $\|\beta\|_1$. $\hat{\beta}$, $\hat{z}$ are functions of $(S, \omega)$. We formally define the optimization map as follows:

$$(s, \omega) \mapsto (s, \hat{\beta}(s, \omega), \hat{z}(s, \omega)) \in S^F(\ell),$$

where

$S^F(\ell) \overset{\text{def}}{=} \left\{ (s, \beta, z) : s \in \text{supp}(F), \ell(\beta; s) < \infty, \|\beta\|_1 < \infty, z \in \partial \|\beta\|_1 \right\}$. (5)

$S^F(\ell)$ is the set of possible values $(s, \beta, z)$ where there will be a solution to (1). We call $S^F(\ell)$ the augmented parameter space. Note $\hat{\beta}$ and $\hat{z}$ are random variables (through the randomness in $(S, \omega)$). One way to describe their distribution, is to find the inverse of the map $\hat{\theta}$ to reconstruct $\omega$ from $(S, \hat{\beta}, \hat{z})$. In the following passage, we denote $\hat{\beta}$, $\hat{z}$ as the random variables and $\beta$, $z$ as the corresponding integration variables when writing out the density.
2.3. Reconstruction and description of the constrained set

Let $\hat{\beta}_E$, $\hat{\omega}$ restricted to $E$, and $\hat{\omega}_E$ the subgradients restricted to $E^c$. To make the notation easier, we define the gradient map $\gamma$

$$\gamma(s, \beta) = \frac{\partial}{\partial \beta} \ell(\beta; s).$$

**Lemma 2.** Through the reparametrization in the optimization map (4), the selection event $\{ E(S, \omega) = E \}$ is equivalent to

$$\begin{aligned}
\gamma(S, \hat{\beta}) + \lambda \cdot \hat{\omega} = 0, \\
\text{sign}(\hat{\beta}_E) = \hat{\omega}_E, \quad \|\hat{\omega}_E\|_\infty \leq 1.
\end{aligned}$$

Lemma 2 provides a reconstruction of $\omega$ using $S$, $\hat{\beta}$ and $\hat{\omega}$. The reconstruction map is defined as,

$$\psi(s, \beta, z) \overset{\text{def}}{=} (s, \gamma(s, \beta) + \lambda \cdot z) = (s, \omega).$$

It is thus easy to see that the distribution of $(S, \hat{\beta}, \hat{\omega})$ follows satisfies the following distributional condition,

$$(S, \gamma(S, \hat{\beta}) + \lambda \cdot \hat{\omega}) \sim F \times G.$$

Moreover, using Lemma 2, the distribution for inference (2) can be rewritten as

$$S \mid (\hat{\beta}(S, \omega), \hat{\omega}(S, \omega)) \in \mathcal{B}, \quad (S, \omega) \sim F \times G,$$

$$\mathcal{B} = \left\{ \hat{\beta}_E = 0, \quad \text{sign}(\hat{\beta}_E) = \hat{\omega}_E, \quad \|\hat{\omega}_E\|_\infty < 1 \right\}.$$

Note that $\mathcal{B}$ is a much nicer set than $\mathcal{A}$ in that it only requires $\hat{\beta}_E$ to be in a certain quadrant and $\|\hat{\omega}_E\|_\infty < 1$.

Combining these two observations above, we have the following theorem. We denote by $T_{E \setminus j} \in \mathbb{R}^{n-1}$ the sufficient statistics $T \in \mathbb{R}^p$ restricted to the set $E \setminus \{j\}$.

**Theorem 3 (Sampling for MAGIC).** Through change of variables (4), the law for selective inference (2) is equivalent to

$$(S, \hat{\beta}(S, \omega), \hat{\omega}(S, \omega)) \in \mathcal{B}, \quad (S, \omega) \sim F \times G, \quad (\hat{\beta}(S, \omega), \hat{\omega}(S, \omega)) \in \mathcal{B}.$$

Moreover, suppose $F$, $G$ has densities $f$ and $g$ respectively, the joint distribution of $(S, \hat{\beta}, \hat{\omega})$ has density proportional to

$$f(s) \cdot g(\gamma(s, \beta) + \lambda \cdot z) \cdot J\psi(s, \beta, z) \cdot 1_B(\beta, z)$$

with the Jacobian denoting the determinant of the derivative of the map $\psi$ with respect to $(\beta, z)$ on the fiber over $s$.

Furthermore, assuming the assumptions in Theorem 1, $P_j(S)$ can be computed (with approximation) with samples from (7) and further conditional on the sufficient statistics $T_{E \setminus j}(S)$. 
Theorem 3 gives the explicit density of the law (7) up to a constant. In the proof we specify how to use the samples from (7) to approximate the p-value function \( P_j \). A natural choice of sampling would be the Metropolis-Hastings method or perhaps the projected Langevin method [3]. To condition on \( T_E \), we just need to make sure the proposal does not move \( T_E \) in each step. Such choice of the proposal is usually natural, for examples see Section 6. The boundary constraint is \( B \), which needs small adjustment from the original Metropolis-Hasting method. Detailed description is in Section 6. After acquiring such samples, we can use them to approximate the p-value function \( P_j \) in Theorem 1.

All the previous work on selective inference also conditions on the observed signs \( \hat{z}_E \).

\[
S \mid (S, \omega) \in \mathcal{A}, \quad \text{sign} \left( \hat{\beta}_E(S, \omega) \right) = \hat{z}_{E, \text{obs}}
\]

where \( \hat{z}_{E, \text{obs}} = \text{sign}(\hat{\beta}(S_{\text{obs}}, \omega_{\text{obs}})) \) is considered fixed. The work [6] explains that any inference valid under this law, would be valid under (2). Note the additional constraint simply requires \( \hat{\beta}_E \) to be in the quadrant specified by \( \hat{z}_{E, \text{obs}} \). In what follows, we also condition on \( \hat{z}_E \).

3. Examples

3.1. Randomized Lasso

Consider linear regression setting where data \( y \sim N(Xb, \sigma^2 I) \), \( X \in \mathbb{R}^{n \times p} \) is fixed, \( \sigma^2 \) is known. Instead of solving the original Lasso proposed by [16], we solve the following randomized version of it,

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \| y - X\beta \|^2_2 + \lambda \| \beta \|_1 - \omega^T \beta.
\] (9)

The gradient of the loss \( \gamma(y, \hat{\beta}) = -X^T(y - X\hat{\beta}) \) and \( \hat{z} \) is the subgradient for \( \| \hat{\beta} \|_1 \). The reconstruction map \( \psi(y, \beta, z) = (y, \lambda z - X^T(y - X\beta)) \). Suppose \( E \) is the active set of (9), then we model the data by \( \mathcal{F} = N(X_E b_E, \sigma^2 I), S = y \) and \( b_E \) is the target for inference.

**Corollary 4** (Randomized Lasso sampler). If \( E \supseteq \text{supp}(b) \), then conditioning on \( (E, \hat{z}_{E, \text{obs}}) \), the joint distribution of \( (y, \hat{\beta}_E, \hat{z}_{-E}) \) can be used for inference (for \( b_E \)). Further, it has density proportional to

\[
\exp \left( -\frac{\| y - X_E b_E \|^2}{2\sigma^2} \right) \cdot g \left( \lambda \left( \frac{z_{E, \text{obs}}}{z_{-E}} \right) - X^T(y - X_E \hat{\beta}_E) \right)
\] (10)

supported on \( \text{sign}(\beta_E) = \hat{z}_{E, \text{obs}} \) and \( \| z_{-E} \|_\infty < 1 \).

We thus can obtain samples \( (y, \hat{\beta}_E, \hat{z}_{-E}) \) for any \( b_E \) in the null hypothesis and use Theorem 1 and Theorem 3 to construct valid p-values. Detailed algorithm is specified in Section 6.
3.2. *L*_1-penalized logistic regression

In practice, many statistical learning problems are classification problems, e.g. spam classification, tumor analysis, etc. Suppose $x_i \sim F_X, x_i \in \mathbb{R}^p, y_i | x_i \sim \text{Bernoulli}(x_i^T b), F_X$ is unknown and $p$ fixed, $S = (X, y)$. The logistic loss is

$$\ell(\beta) = -\frac{1}{\sqrt{n}} \left[ \sum_{i=1}^{n} y_i \log \pi(x_i^T \beta) + (1 - y_i) \log(1 - \pi(x_i^T \beta)) \right],$$

where $\pi(x) = \exp(x) / (1 + \exp(x))$. The randomized logistic regression solves the following problem

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \ell(\beta) + \lambda \|\beta\|_1 - \omega^T \beta + \frac{\epsilon}{2} \|\beta\|^2_2 \tag{11}$$

with $\epsilon > 0$ small and fixed. The addition of the term with $\epsilon$ is to ensure the existence of the solution to (11). We explicitly express the $\epsilon$ term, but in general it can be absorbed into the loss function.

Suppose $E$ is the active set of (11), then $b_E$ is the target of inference. With slight abuse of notation, we allow $\pi : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto \pi(x)$ to be the function applied on each coordinate of $x \in \mathbb{R}^n$. With some algebra, we have the reconstruction map for $\omega$

$$\omega = \lambda \cdot \hat{z} - \frac{1}{\sqrt{n}} X^T \left[ y - \pi(X \hat{\beta}) \right] + \epsilon \hat{\beta}$$

To sample $(X, y)$ jointly is not feasible when $F_X$ is unknown. Two observations help us circumvent it and even make the sampling more efficient. First, the reconstruction map for $\omega$ only involve the random vector

$$\nabla \ell(\hat{\beta}_E) = -\frac{1}{\sqrt{n}} X^T (y - \pi(X \hat{\beta}_E))$$

$$\approx -\frac{1}{\sqrt{n}} X^T \left( y - \pi(X \hat{\beta}_E) - W(X \hat{\beta}_E)X \hat{\beta}_E - \hat{\beta}_E \right)$$

where $\hat{\beta}_E$ is the MLE for the unpenalized logistic regression with only the variables in $E$ and $W(X \beta) = \text{diag}(\pi(X \beta) (1 - \pi(X \beta)))$ is the weight matrix. Alternatively, we might take $\hat{\beta}_E$ to be the one-step estimator in the selected model starting from $\hat{\beta}_E$ [13]. The gradient $\nabla \ell(\hat{\beta}_E)$ can be reconstructed, up to a Taylor remainder, from $\hat{\beta}$ and the random vector

$$T = \left( X^T_E (y - \pi(X \hat{\beta}_E)) \right).$$

Moreover, when $p$ is fixed, pre-selection, our random vector $T$ properly scaled is asymptotically normal and when the selected model is correct ($E \supseteq \text{supp}(b)$):

$$\frac{1}{\sqrt{n}} \left[ T - \left( \begin{array}{c} b_E \\ 0 \end{array} \right) \right] \xrightarrow{D} N(0, \Sigma) \tag{12}$$

where $\Sigma$ is estimable from the data [13]. Since asymptotically $T$ is from an exponential family with parameters $b_E$, Theorem 1 states the p-value is a function of $T$ only. Thus instead of sampling $(X, y)$, we only need to sample the distribution $T$. 
Theorem 5. Suppose $E \supseteq \text{supp}(b)$ and conditioning on $(E, z_{E, \text{obs}})$, the joint distribution of $(T, \hat{\beta}_E, \hat{z}_E)$ can be used for inference. Then the distribution of $(T, \hat{\beta}_E, \hat{z}_E)$ asymptotically (with $p$ fixed, $n \to \infty$) has density

$$\phi(T) \cdot g \left( \frac{1}{\sqrt{n}} \left( X^T_E W (X_E \hat{\beta}_E) X_E (\hat{\beta}_E - \beta_E) + X^T_E (y - \pi (X_E \hat{\beta}_E)) \right) + \lambda \begin{pmatrix} z_{E, \text{obs}} \\ z_{-E} \end{pmatrix} + \epsilon \begin{pmatrix} \hat{\beta}_E \\ 0 \end{pmatrix} \right)$$

(13)

where $\phi$ is the density for $N((b_E, 0), \Sigma)$.

3.3. Neighborhood Selection

Gaussian graphical models have recently become a very popular way to study network structures. In particular, it has often been used on many types of genome data (e.g. gene expression, metabolite concentrations etc.) Suppose the data we observe is $X \in \mathbb{R}^{n \times p}$, where each row of $X$ is independently distributed as $N(\mu, \Sigma), \mu \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}$.

It is of interest to study the conditional independence structure of the variables $\{1, 2, \ldots, p\}$. The conditional independence structure is conveniently represented by an undirectional graph $(\Gamma, E)$, where the nodes $\Gamma = \{1, 2, \ldots, p\}$, and there is an edge between $i, j$ if and only if $x_i \perp x_j$ conditional on all the other variables $\Gamma \setminus \{i, j\}$. Moreover, assuming the covariance matrix $\Sigma$ is not singular, we denote the inverse covariance matrix $\Theta = \Sigma^{-1}$, then

$$x_i \perp x_j | X_{\Gamma \setminus \{i, j\}} \iff \Theta_{ij} = 0.$$  

In many applications of Gaussian graphical models, we assume the sparse edge structure, where we can hope to recover the edgset $E$ even when $n < p^2$. The authors in [8] proposed neighborhood selection with the Lasso to achieve this goal. The algorithm can be formulated as the following optimization problem, for any node $i$

$$\hat{\beta}_i^{1, \lambda} = \arg\min_{\beta_i, \beta_i=0} (n^{-1} \| x_i - X \beta \|_2^2 + \lambda \| \beta \|_1), \tag{14}$$

where $x_i$ is the $i$-th column of $X$, $\lambda$ is chosen according to Chapter 3 of [8] and considered fixed. Denote $\hat{B} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$, we propose the randomized version of (14),

$$\hat{B} = \arg\min_{B : B_{ii}=0} \| X - X B \|_F^2 + \lambda \| B \|_1 - \Omega B, \tag{15}$$

where $\Omega = (\omega^1, \ldots, \omega^p)$, $\omega^i = \chi_i^d G$. Let $E' = \text{supp}(\hat{\beta}^i)$, and $E = (E^1, \ldots, E^p)$. Since $E$ is usually not symmetric, we instead look at the set

$$E'^\vee = \{(i, j) | E_{ij} = 1 \text{ or } E_{ji} = 1\}.$$  

Our target for inference is $\{\Theta_{ij}, (i, j) \in E'^\vee\}$. Note (15) is the matrix form of (1), and the reconstruction maps are decomposable across the $p$ nodes; Therefore, we have the following corollary,
Corollary 6. Suppose $E$ is the active set for (15), and $\hat{z}_E$ is the corresponding signs of $\hat{B}_E$, then conditioning on $(E, z_E, obs)$, the distribution of $(X, \hat{B}_E, \hat{z}_E)$ can be used for inference. Furthermore, if we assume $\Theta_{ij} = 0, \ i \neq j$ and $(i, j) \notin E^\vee$, then the joint distribution of $(X, \hat{B}_E, \hat{z}_E)$ has the following density,

$$
\exp \left[ -\frac{1}{2} \sum_{i=1}^{p} \Theta_{ii} \|x_i\|^2 + \sum_{(i,j) \in E^\vee} \Theta_{ij} x_i^T x_j \right] \cdot \prod_{i \in \Gamma} g \left( \begin{pmatrix} z_{E, i, obs} \\ z_{E, i}^{\vee} \end{pmatrix} - X_{-i}^T (x_i - X_{E, i} \beta_{E, i}) \right) \cdot \det(X_{E, i}^T X_{E, i}).
$$

4. Simulation

Theorem 1 states that our p-values should be valid at level $\alpha$, for any $\alpha \in [0, 1]$, see (3). In fact, all the three examples above satisfy the condition such that the Type-I error for any level-$\alpha$ test would be equal to (or asymptotically equal to) $\alpha$. That is equivalent as saying the p-values follow Unif$(0, 1)$ distribution. To validate Theorem 1 and Theorem 3, we ran the following simulations for each of the examples in Section 3. Our data is generated as follows, for Lasso,

$$y \sim N(Xb, \sigma^2 I), \ X \in \mathbb{R}^{n \times p}, \ \text{fixed}, \ \|b\|_0 = s,$$

where $s \ll p$. The framework works for arbitrary $n$ and $p$. To demonstrate the applicability of our framework in high dimensions, we set $n = 50, \ p = 100, \ s = 7$. For logistic Lasso problem,

$$x_i \sim N(0, I), \ y_i | x_i \sim \text{Bernoulli}(\pi(x_i b)), \ \pi = \frac{\exp(x)}{1 + \exp(x)}, \ \|b\|_0 = s.$$

The framework for logistic regression is fixed $p$ and $n \to \infty$. Thus we take $n = 500, \ p = 50, \ s = 5$. For both of the examples above, the signal to noise ratio (snr) is 7. Finally, for neighborhood selection, the data matrix is $X \in \mathbb{R}^{n \times p}$, each row of $X$ is i.i.d from $N(0, \Theta^{-1})$. We take $n = 100, \ p = 30$, note this is a high-dimensional setting since we have $30 \times 30$ unknown parameters. But only 1% of off-diagonal elements of $\Theta$ is non-zero, and the non-zero off-diagonal entries of $\Theta$ are taken to be $\rho = 0.245$ and the diagonal elements are 1. $\rho = 0.245$ is chosen because any value less than 0.25 would ensure $\Theta$ is positive definite [8].

For each $j \in E$, we test the hypothesis $H_{0j} : b_j = 0$, against a two-sided alternative hypothesis. We call the p-values the null p-values when the null hypothesis is true and alternative p-values otherwise. When the active set $E$ (or $E^\vee$) from the problem covers $\text{supp}(b)$ (or $\text{supp}(\Theta)$), the null p-values should follow Unif$(0, 1)$. Figure 1 is the plot for the empirical cdf for the null p-values computed from Lasso, logistic Lasso and neighborhood selection. We see that all the null-pvalues follow the uniform distribution, verifying our Theorem 1 and Theorem 3.
4.1. Comparisons of statistical powers

As we mentioned in Section 1, randomization significantly boosts power. This is shown in both hypothesis testing and confidence intervals. We describe what it means in both aspects. For a valid selective level-α test, Type-I error is controlled at α conditional on selection. We hope to achieve valid tests with high power. In the selective inference framework, statistical power is simply defined as the power in the selected model [5, 15]. If $E \supseteq \text{supp}(b)$, then for any $j \in E$,

$$\text{power} = \mathbb{P}[\text{reject } H_{0j} \mid H_{1j} \text{ is true, } E \text{ is selected}].$$

The selective inference framework also offers confidence intervals by inverting a valid test, for examples, see [6]. We want short confidence intervals which have the desired coverage guarantees. MAGIC enjoys higher statistical power (shorter intervals), the tradeoff is slightly worse selected models as we added randomization for model selection. However, the tradeoff is highly in favor of MAGIC. Usually just a small amount of randomization will dramatically increase statistical power. In the linear regression case, this has been shown in [15] with simulated data. In the following passage, we give numerical comparisons on both a real dataset and simulated data.

4.1.1. In vitro HIV drug resistance

In [12], the authors study the genetic basis of drug resistance in HIV, using markers of inhibitor mutations to predict a quantitative measurement of susceptibility to several antiretroviral drugs. The hope is to find the mutations highly correlated with the susceptibility to drugs. We apply Lasso to the protease inhibitor subset of their data and select the potential mutations set for one of the drugs, Lamivudine (3TC). We then compute the OLS estimator in the selected set of gene mutations, and form confidence intervals for the coefficients (Figure 2). The grey bars are the OLS estimates with only the selected mutations. The confidence intervals are adjusted for selection and should have the desired coverage 90%. We report the estimators together with the confidence intervals. The procedure in left panel 2a is the same as [6] without randomization in
Fig 2: Confidence intervals for selected genes in 3TC DATA

(a) Selective intervals without randomization
(b) Selective intervals with randomization

selecting the mutations. The right panel 2b in contrast uses the MAGIC framework for LASSO with randomization $\omega \sim N(0, 0.1\sigma^2)$, where $\sigma^2_{cv}$ is the noise level estimated by cross-validation. Note the mutations selected by the two methods only differ by 3 mutations, with small effects, and the OLS estimator for the common mutations are very close. But the randomized selection procedure gives much shorter confidence interval across all mutations, demonstrating the advantage and practicality of our methods.

4.1.2. Statistical power comparison with simulated data

In this section, we compare more specifically the tradeoff between power and model selection using simulated data. The authors in [13] offered explicit calculations of p-values after the model is selected by $\ell_1$ penalized logistic regression or graphical Lasso. Both examples can be considered in the MAGIC framework. Simulations in [13] showed that graphical Lasso has worse power than $\ell_1$ penalized logistic regression. Therefore, we compare our framework to the latter. We assume the same setup as before, our randomization noise is $\omega \sim N(0, 0.1\sigma^2)$ and $\epsilon = 0.02$. The proportions of selecting the “true” models ($E \supseteq \text{supp}(b)$) is 0.91 without randomization and 0.852 in MAGIC. Much more different is the power of the two procedures; for a level-0.05 test, the statistical powers defined above is 0.176 without randomization and 0.887 in MAGIC. Figure 3 is the histograms for the alternative p-values with or without randomization.

5. Proofs

5.1. Proof for Theorem 1

Proof. Let $S$ be the space for $S$, then $(S, \omega) \in S \times \mathbb{R}^p$. The joint distribution of $(S, \omega)$ conditional on $(S, \omega) \in A$ has the following density with respect to the measure $\mu(ds)G(d\omega)$

$$h(s, \omega) = \frac{\exp[b^T T(s)] \mathbf{1}\{(s, \omega) \in A\}}{\int_{S \times \mathbb{R}^p} \exp[b^T T(s)] \mathbf{1}\{(s, \omega) \in A\} \mu(ds)G(d\omega)}.$$
Fig 3: The alternative p-values computed from the MAGIC framework highly concentrated around 0, while without randomization the p-values are more evenly distributed between [0, 1]. The statistical powers are 0.887 for MAGIC vs. 0.176 for non-randomized procedure with a level-0.05 test.

Since the denominator is merely a normalizing constant, (17) is also an exponential distribution with parameters $b$, sufficient statistics $T(S)$ and a slightly different reference measure $\mathbf{1}\{(s, \omega) \in A\} \mu(ds)G(d\omega)$. Since $E \supseteq \text{supp}(b)$, $b^T T(s) = b_E^T T_E(s)$, where $b_E, T_E \in \mathbb{R}^{|E|}$ are $b$ and $T$ restricted to set $E$. Thus (17) can be seen as an exponential family with sufficient statistics $T_E$ and parameters $b_E$. To test any hypothesis $H_{0j} : b_j = 0$, $j \in E$, Chapter 4 of [7] states that Uniformly Most Powerful Unbiased tests can be constructed using the statistic $T_j$ and conditioning on all the other sufficient statistics $T_{E \setminus j} \in \mathbb{R}^{|E|-1}$. Thus the conditional density of the one dimensional distribution for $T_j$ is

$$h_j(t_j; t_{E \setminus j}) = \frac{\exp \left[ b_j t_j + b_{E \setminus j}^T T_{E \setminus j} \right] \cdot \int_A \mathbf{1}\{T_{E \setminus j}(s) = t_{E \setminus j}\} \mu(ds)G(d\omega)}{\int_A \exp \left[ b_{E \setminus j}^T T_{E \setminus j} \right] \cdot \mathbf{1}\{T_{E \setminus j}(s) = t_{E \setminus j}\} \mu(ds)G(d\omega)}$$

Thus (18) is the density for the distribution

$$T_j(S) | T_{E \setminus j}(S), (S, \omega) \in A, (S, \omega) \sim F \times G.$$  \hspace{1cm} (19)

Note (18) involves only the parameter $b_j$, thus it can be used to test the composite hypothesis $H_{0j} : b_j = 0$, with $b_{E \setminus j}$ taking arbitrary values.

Let $H_j$ denote the c.d.f of the above law: $H_j(t_j; T_{E \setminus j}) = \int_{-\infty}^{t_j} h_j(r; T_{E \setminus j}) \, dr$. Then we can construct our function $\tilde{P}_j : \mathbb{R}^p \to \mathbb{R}$ as

$$\tilde{P}_j(t) = \frac{\int_A \exp \left[ b_j T_j(s) \right] \mathbf{1}\{T_j(s) > t_j\} \mathbf{1}\{T_{E \setminus j}(s) = t_{E \setminus j}\} \mu(ds)G(d\omega)}{\int_A \exp \left[ b_j T_j(s) \right] \mathbf{1}\{T_{E \setminus j}(s) = t_{E \setminus j}\} \mu(ds)G(d\omega)}$$

$$= 1 - H_j(t_j; t_{E \setminus j}).$$  \hspace{1cm} (20)
Under the null hypothesis, we take $b_j = 0$, thus $\tilde{P}_j$ depends only on $T(s)$ and $A$. We define $P_j = \tilde{P}_j(T(s))$. Now we prove the level-$\alpha$ control (3). Note

$$P_F \times G[P_j(S) \leq \alpha | (S, \omega) \in A] = E[P_F \times G[P_j(S) \leq \alpha | T_{E \setminus j}, (S, \omega) \in A],$$

it suffices to prove the quantity inside the expectation has the level-$\alpha$ control for any $T_{E \setminus j}$. Since $H_j$ is the c.d.f of the conditional law (19),

$$P_F \times G[P_j(S) \leq \alpha | T_{E \setminus j}, (S, \omega) \in A] = P_F \times G[1 - H_j^{-1}(1 - \alpha) | T_{E \setminus j}, (S, \omega) \in A] = 1 - H_j^{-1}(1 - \alpha) \leq \alpha,$$

where $H_j^{-1}$ generalized inverse for $H_j$, the equality holds when $H_j$ is strictly increasing in $t_j$.

5.2. Proof for Lemma 2

**Proof.** Equation (1) is a convex optimization problem. The solution $\hat{\beta}$ and subgradient of the $\ell_1$ norm $\hat{z}$ satisfy the KarushKuhnTucker conditions (KKT), which are sufficient and necessary.

$$\begin{cases}
\partial \beta \ell(\hat{\beta}; S) + \hat{z} - \omega = 0, \\
\hat{z} \in \partial \|\hat{\beta}\|_1.
\end{cases}$$

The equations are simply the differentiation of the optimization objective function. This gives the equation part in (6) of the lemma. Note the penalty term $\|\hat{\beta}\|_1$ is differentiable except at 0, its subgradient at 0 is $[-1, 1]$. Thus, conditioning on the active set $\hat{E}(S, \omega) = E$ it is equivalent to:

$$\begin{cases}
\hat{z}_j = \text{sign}(\hat{\beta}_j), \quad \forall j \in E, \\
|\hat{z}_j| \leq 1, \quad \forall j \notin E.
\end{cases}$$

Combining the above two, we have the conclusion of the lemma.

5.3. Proof for Theorem 3

**Proof.** Per the discussion above Theorem 3, it is not hard to see the distributional constraint on $(S, \hat{\beta}, \hat{z})$ is that $\gamma(S, \hat{\beta}) + \lambda \hat{z} \sim G$ and is independent of $S$. Moreover, $(\hat{\beta}, \hat{z})$ are constrained to be in the region $\hat{B}$. Thus the law (7) is the marginal law of $S$ conditional on selection.

Now we investigate the joint density of $(s, \beta, z)$. Through the reconstruction map,

$$\psi(s, \beta, z) = (s, \gamma(s, \beta) + \lambda z),$$
the density of \((s, \beta, z)\) is simply the product \(f(s)g(\gamma(s, \beta) + \lambda z)\) times the determinant of the Jacobian matrix. Standard multivariate calculus yields the form of the Jacobian matrix of \(\psi\) as follows,

\[
\begin{pmatrix}
I & 0 \\
D_s\psi(s, \beta, z) & D_{(\beta, z)}\psi(s, \beta, z)
\end{pmatrix}
\]

with determinant \(\det D_{(\beta, z)}\psi(s, \beta, z)\). Thus we have (8).

Notice the construction (20), \(P_j(s) = 1 - H_j(T_j(s), T_{E \setminus j}(s))\) and \(H_j\) is the c.d.f for the conditional distribution (19). It is equivalent to sampling (7) while further conditional on \(T_{E \setminus j}(S)\). After we acquire \(m\) samples \(\{S(1), \ldots, S(m)\}\), we can approximate the integral in (20) as the percentile of \(T_j(S)\) among \(\{T_j(S(1)), \ldots, T_j(S(m))\}\). \(\square\)

5.4. Proof for Corollary 4

Proof. Notice that once we condition on the active set \(E\), and the signs \(z_{E, \text{obs}}\), \(\beta = (\beta_E^0)\), \(z = (z_{E, \text{obs}}, z_{-E})\). (21)

Therefore, the density of \((y, \beta, z)\) is equivalent to that of \((y, \beta_E, z_{-E})\), through the construction in (21). Note the Jacobian matrix is

\[
D_{\beta_E, z_{-E}}\psi(y, \beta_E, z_{-E}) = \begin{pmatrix}
X_E^TX_E & 0 \\
X_E^TX_E & \lambda I
\end{pmatrix},
\]

where \(I\) is the identity matrix of dimension \(p - |E|\). Then the Jacobian \(J\psi(y, \beta, z) = \lambda^{p-|E|}\det(X_E^TX_E)\). Since the Jacobian is a constant only depending on \(X\). Thus, if we plug in \(\beta, z\) in Theorem 3, the density of \((y, \beta_E, z_{-E})\) has the form in Corollary 4. \(\square\)

5.5. Proof for Theorem 5

Proof. We first reconstruct the gradient \(\nabla \ell(\hat{\beta}_E)\) from \(\hat{\beta}_E\) and

\[
T = \begin{pmatrix}
X_E^T(y - \pi(X_E\hat{\beta}_E))
\end{pmatrix}.
\]

The Taylor expansion of \(\nabla \ell(\tilde{\beta}_E)\) at \(\tilde{\beta}_E\) is

\[
\nabla \ell(\tilde{\beta}_E) = -\frac{1}{\sqrt{n}}X^T(y - \pi(X_E\tilde{\beta}_E)) = -\frac{1}{\sqrt{n}}X^T\left(y - \pi(X_E\tilde{\beta}_E) - W(X_E\tilde{\beta}_E)X_E(\hat{\beta}_E - \tilde{\beta}_E)\right) + R,
\]

where \(R = o_p(1)\). Since \(\hat{\beta}_E\) is the minimizer of the logistic regression with \(E\) variables, the gradient at \(\hat{\beta}_E\) is zero,

\[
X_E^T(y - \pi(X_E\hat{\beta}_E)) = 0.
\]
Thus we can rewrite $\nabla \ell(\hat{\beta}_E)$ in terms of $\hat{\beta}_E$ and $T$ via the following map,

$$\nabla \ell(\hat{\beta}_E) = \frac{1}{\sqrt{n}} \left( X_{-E}^T W(X_E \hat{\beta}_E) X_E (\hat{\beta}_E - \hat{\beta}_E) \right) + R. \quad (22)$$

Notice that $\hat{\beta}_E$ is the MLE for the negative logistic likelihood, and thus satisfy the asymptotic normality, with asymptotic mean $b_E$, when $E \supseteq \text{supp}(b)$. Moreover, the following part has asymptotically mean 0,

$$\mathbb{E} \left[ \frac{1}{\sqrt{n}} X_{-E}^T y - \pi(X_E \hat{\beta}_E) \right] - \mathbb{E} \left[ \frac{1}{\sqrt{n}} X_{-E}^T W(X_E b_E) X_E (\hat{\beta}_E - b_E) \right] + o_p(1) \to 0.$$

Thus we have the asymptotic normality as in (12). Moreover, since $\omega = \nabla \ell(\hat{\beta}_E) + \lambda \hat{z} + \epsilon \hat{\beta}_E$, then we have asymptotically,

$$(T, \nabla \ell(\hat{\beta}_E) + \lambda \hat{z} + \epsilon \hat{\beta}_E) \overset{d}{\to} F \times G.$$  

The Jacobian is $\det(X_{-E}^T W(X_E \hat{\beta}_E) X_E + \epsilon I)$ which by law of large numbers converges to $\det \left( \mathbb{E}(X_{-E}^T W(X_E b_E) X_E) + \epsilon I \right)$, a constant. Therefore, we have the density (13) if we plug in the map (22) for $\nabla \ell(\hat{\beta})$.  

5.6. Proof for Corollary 6

Proof. For every node $i$, the $i$-th coordinate of $\beta^i$ is held to be zero, and (14) is in fact a regression of dimension $p - 1$, thus $\gamma(X, \hat{\beta}^i) = -X_{-i}^T (x_i - X \hat{\beta}^i) \in \mathbb{R}^{p-1}$, and the reconstruction map,

$$\psi : (X, \hat{B}, \hat{z}) \mapsto (X, \gamma(X, \hat{B}) + \lambda \hat{z}),$$

where

$$\gamma(X, \hat{B}) = (\gamma(X, \hat{\beta}^1), \ldots, \gamma(X, \hat{\beta}^p)) \in \mathbb{R}^{(p-1) \times p},$$

$$\hat{z} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_p), \hat{B} = (\hat{\beta}^1, \ldots, \hat{\beta}^p),$$

and $\hat{z}_i = \begin{pmatrix} z_{1E, \text{obs}}^i \\ z_{\hat{E}, E}^i \end{pmatrix}$ is the subgradient of the optimization problem (14). Since $\omega^i$'s are independent, and the Jacobian

$$J \psi(X, z, B) = \prod_{i \in \Gamma} \det(X_{E_i}^T X_{E_i}'),$$

density (16) follows.
Algorithm 1 Metropolis Hastings sampler for randomized Lasso

Set: $b = 0$ for distribution $f_b$, compute the explicit expression $h$.
Compute: $P = X_E \setminus X_E^T, R = I - P$.
Initialize: $(y_0, \hat{\beta}_E, \hat{\omega}_E, \hat{\tau}_E) \leftarrow (y, \hat{\beta}_E, \hat{\omega}_E)$.

Step data: $y^{(n+1)} \leftarrow Py^{(n)} + a_n \cdot R \tau, \ \tau \sim N(0, I)$, compute the acceptance ratio $r = \frac{h(y^{(n+1)}, \hat{\beta}_E^{(n)}, \hat{\omega}_E^{(n)})}{h(y^{(n)}, \hat{\beta}_E^{(n)}, \hat{\omega}_E^{(n)})}$, accept $y^{(n+1)}$ with probability $r$, otherwise $y^{(n+1)} \leftarrow y^{(n)}$. If $r > 1$, accept $y^{(n+1)}$. 

Step coefficient: $\hat{\beta}_E^{(n+1)} \leftarrow sE|\hat{\beta}_E^{(n)} + c_n \cdot \nu|, \ \nu \sim G$, compute the acceptance ratio $r = \frac{g(y^{(n+1)}, \hat{\beta}_E^{(n+1)}, \hat{\omega}_E^{(n)})}{g(y^{(n+1)}, \hat{\beta}_E^{(n)}, \hat{\omega}_E^{(n)})}$, and accept/reject accordingly.

Step subgradient: compute the upper and lower limits,
\[
\Delta^+ = -X_E^T(y^{(n+1)} - X_E\hat{\beta}_E^{(n+1)}) + \lambda I,
\]
\[
\Delta^- = -X_E^T(y^{(n+1)} - X_E\hat{\beta}_E^{(n+1)}) - \lambda I,
\]
sample $\hat{\omega}_E^{(n+1)}$, $\hat{\tau}_E^{(n+1)} \sim G_{\Delta^-, \Delta^+}$.

6. Monte-Carlo sampler

Theorem 3 gives an explicit way of computing the density for the law of selective inference. We can use a Gibbs sampler to rotate through sampling $(S, \hat{\beta}, \hat{\omega})$. For sampling $S$ and $\hat{\beta}$, we can take a Metropolis-Hastings step and use the density to compute the acceptance probability. For sampling $\hat{\omega}$, it is even simpler as we recognize the conditional distribution of $\hat{\omega}|S, \hat{\beta}$ simply a truncated $G$ distribution. To illustrate our sampler, we describe the sampling scheme of some of our examples in more details.

6.1. Randomized Lasso sampler

Without loss of generality, we assume the density of added noise $G$ is symmetric and each coordinate of $\omega$ is independent. This is in fact what we use a lot in practice. Also denote $G_{\Delta^-, \Delta^+}$ as truncated distribution $G$ with $\Delta^-$, $\Delta^+$ as the lower and upper truncation points, and $h(y, \hat{\beta}_E, \hat{\omega}_E)$ to be the density in (10). Then to test the null hypothesis $H_{0ij} : b_1 = 0$, we propose Algorithm 1. Note the step sizes $a_n$ and $c_n$ in Algorithm 1 is chosen through [18] to achieve the desired acceptance rate.

6.2. Neighborhood selection

Similar to the scheme in Section 6.1, we use a Gibbs sampler to sample $X, \hat{B}$ and $\hat{\omega}$ respectively. The sampling for $\hat{B}$ and $\hat{\omega}$ are analogous to that of Section 6.1, and we only need a proposal distribution for $X$. As mentioned in Section 3.3, to test the hypothesis $H_{0ij} : \Theta_{ij} = 0$, we condition on $\{x_i^T x_j, (i', j') \neq (i,j)\}$. To sample the data matrix $X$, we rotate through its columns, sampling one column at a time, keeping all the others as constant. More specifically, for column $i$, we sample from the distribution,

$$x_i \mid X_{-i}, \|x_i\|^2, x_i^T x_{j'}, (i', j') \in E^\lor, (i', j') \neq (i,j).$$
Note the graph structure gives a natural partition of the nodes into different connected components, let \( ne(i) \) be the nodes in the connected component of \( i \), then \( x_i \perp x_k, \forall k \notin ne(i) \), conditioning on all the other \( x_j', j \in ne(i) \). Thus the above law is equivalent to,

\[
x_i \|x_i\|^2, x_j, x_i^T x_j, j \in ne(i).
\] (23)

We can sample the above law (23) by sampling uniformly from a sphere with radius \( ||x_i|| \), holding the projections onto the \( x_j \)'s constant. After sampling a new column of \( X \), we compute the accept ratio, accept/reject accordingly and move to the next column. As for the sampling of \( \hat{B} \) and \( \hat{z} \), we can develop an algorithm similar to Algorithm 1.

7. Discussion

MAGIC has the following limitations that we hope to remove in future work. First, the penalty in our convex program have to be \( \ell_1 \) penalty. Second, we assume parametric models, more specifically in the exponential family setting. Third, in the setting for Section 3.2, we require the dimension \( p \) to be fixed, leaving the high-dimensional problem \( p > n \) as an interesting problem.

References

[1] Claude JP Bélisle, H Edwin Romeijn, and Robert L Smith. Hit-and-run algorithms for generating multivariate distributions. Mathematics of Operations Research, 18(2):255–266, 1993.
[2] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. The Annals of Statistics, 41(2):802–837, April 2013.
[3] S. Bubeck, R. Eldan, and J. Lehec. Sampling from a log-concave distribution with Projected Langevin Monte Carlo. ArXiv e-prints, July 2015.
[4] DR Cox. A note on data-splitting for the evaluation of significance levels. Biometrika, 62(2):441–444, 1975.
[5] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. arXiv:1410.2597 [math, stat], October 2014. arXiv: 1410.2597.
[6] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference with the lasso. arXiv:1311.6238 [math, stat], November 2013.
[7] Erich L. Lehman. Testing Statistical Hypotheses. Springer-Verlag, New York, 2nd edition, 1997.
[8] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, pages 1436–1462, 2006.
[9] Nicolai Meinshausen and Peter Bühlmann. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4):417–473, 2010.
[10] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. Journal of the American Statistical Association, 2012.
[11] Sahand Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of $\ell_1$-estimators with decomposable regularizers. arXiv:1010.2731, October 2010.
[12] Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, Douglas L Brutlag, and Robert W Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006.

[13] Jonathan Taylor and Robert Tibshirani. Post-selection inference for l1-penalized likelihood models. *arXiv preprint arXiv:1602.07358*, 2016.

[14] Xiaoying Tian, Joshua R. Loftus, and Jonathan E. Taylor. Selective inference with unknown variance via the square-root LASSO. *arXiv:1504.08031 [math, stat]*, April 2015. arXiv: 1504.08031.

[15] Xiaoying Tian and Jonathan E. Taylor. Selective inference with a randomized response. *arXiv:1507.06739 [math, stat]*, July 2015. arXiv: 1507.06739.

[16] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[17] Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pages 614–645, 2008.

[18] Matti Vihola. Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008, 2012.

[19] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.

[20] Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

[21] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

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