Selective inference with unknown variance via the square-root LASSO

Xiaoying Tian, and Joshua R. Loftus, and Jonathan E. Taylor

Department of Statistics
Stanford University
Stanford, California
e-mail: xtian@stanford.edu, joftius@stanford.edu

Abstract: There has been much recent work on inference after model selection when the noise level is known, for example in forward stepwise model selection or LASSO with an independent estimate of $\sigma$. In this work we consider the more realistic scenario of an unknown noise level and propose using the square root LASSO (also known as the scaled LASSO) to conduct selective inference without previous knowledge of $\sigma$. Applying the selective inference framework described in Fithian et al. (2014), we construct selectively valid exact tests of parameters in the selected model. We discuss regression diagnostics including tests for inclusion of variables not selected by the square root LASSO. The selective inference framework also suggests a natural pseudo-likelihood estimate of the noise level that performs better than other estimates of $\sigma$ from the square root LASSO. Based on this estimate we use a Gaussian approximation to construct confidence intervals. We also consider exact inference when holding out some data for inference in the second stage, noting the same increase in power observed in Fithian et al. (2014). We illustrate our method on an HIV drug resistance data set and in a multi-scale change point detection problem where the number of regressors is order $n^2$.

AMS 2000 subject classifications: Primary 62F03, 62J07; secondary 62E15.

Keywords and phrases: lasso, square root lasso, confidence interval, hypothesis test, model selection.

1. Introduction

In this work, we consider selective inference Fithian et al. (2014) for a standard regression problem. Selective inference differs from classical inference in regression. Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$ we first choose a model $X_E$ by considering some model selection procedure. Having chosen a model $X_E$, it is natural to consider the canonical regression model

$$y = X_E \beta_E + \epsilon, \epsilon \sim N(0, \sigma^2 E I),$$

(1.1)

and continue on with the usual types of inference considered in regression such as hypothesis tests and construction of confidence intervals.

There has been a fair amount of recent work devoted to this problem Taylor et al. (2013), Lee et al. (2013), Taylor et al. (2014), Fithian et al. (2014). Most of this work assumes some type of sparsity—that there exists a subset $E \subset \{1, \ldots, p\}$ such that $X_E = X[\cdot, E]$ is a good regression model (though not necessarily that it is the true model). In most of this work the variance $\sigma^2 E$, which is crucial for inference, is assumed to be known. For some selection procedures such as the LASSO at a fixed value of the regularization parameter Lee et al. (2013) knowledge of $\sigma$ is also important. Fithian et al. (2014) allows for inference with unknown $\sigma$ but, as the focus was broader than...
regression, did not suggest a particular selection procedure with unknown \( \sigma \). However, assuming a fixed number of steps of forward stepwise or the LAR (Least Angle Regression) algorithm Taylor et al. (2014), or a fixed number of variables in marginal screening Lee & Taylor (2014), then the results in Fithian et al. (2014) could be adapted to carry out inference with unknown \( \sigma \). This would also require moving from the saturated model approach to the selected model approach, a distinction we described in further detail in Section 1.2.

This work focuses on adapting the approach of Lee et al. (2013) to the case of unknown \( \sigma \), both in terms of choice of tuning parameter and post selection inference. It is also a rich and somewhat more realistic application of the selected model framework described in Fithian et al. (2014).

1.1. The Square-root LASSO

The selection procedure we use is based on the square-root LASSO Belloni et al. (2010), which in turn is known to be equivalent to the scaled LASSO Sun & Zhang (2011). The square-root LASSO is a modification of the LASSO Tibshirani (1996):

\[
\hat{\beta}_\gamma = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \| y - X\beta \|^2 + \gamma \cdot \| \beta \|_1. \tag{1.2}
\]

For the LASSO, a choice of \( \gamma \) known to have good estimation properties Negahban et al. (2012) is

\[
\gamma = 2 \cdot \mathbb{E}(\| X^T \epsilon \|_\infty), \quad \epsilon \sim N(0, \sigma_E^2 I) \tag{1.3}
\]

which depends on the noise variance \( \sigma \). In practice, we might consider some multiple other than 2. Decreasing this multiple leads to more false positives selected by the LASSO, though these can be detected by hypothesis tests as in Lee et al. (2013). As \( \lambda_1 = \lambda_1(X, y) \) (the first non-zero solution to (1.2) Lockhart et al. (2013), Efron et al. (2004)) is equal to \( \| X^T y \|_\infty \), the choice of tuning parameter can roughly be viewed as some multiple of the expected threshold at which noise with variance \( \sigma_E^2 \) would enter the LASSO path.

The square-root LASSO is the convex program

\[
\hat{\beta}_\lambda = \arg \min_{\beta \in \mathbb{R}^p} \| y - X\beta \|^2 + \lambda \cdot \| \beta \|_1. \tag{1.4}
\]

Note that we parameterize this slightly differently than in the literature. An analogous choice of \( \lambda \) to the LASSO is

\[
\lambda = \kappa \cdot \mathbb{E} \left( \frac{\| X^T \epsilon \|_\infty}{\| \epsilon \|_2} \right), \quad \epsilon \sim N(0, I) \tag{1.5}
\]

for some unitless \( \kappa \). Below, we typically use \( \kappa \leq 1 \).

Any spherically symmetric distribution yields the same choice of \( \lambda \) which makes the above choice of tuning parameter independent of the noise level \( \sigma \). The independence of \( \lambda \) with noise level \( \sigma \) follows from the convex program (1.4) since both the first term and the second term in the optimization objective contain the same order of \( \sigma \).

Both the LASSO and square-root LASSO can be viewed as model selection procedures. In what follows we make the weak assumption that the columns of \( X \) are in general position Tibshirani (2013). We define the selected model of the square-root LASSO as

\[
\hat{E}_\lambda(y) = \left\{ j : \hat{\beta}_{j,\lambda}(y) \neq 0 \right\} \tag{1.6}
\]
and the selected signs

\[ \hat{z}_{E,\lambda}(y) = \text{sign} \{ \hat{\beta}_{j,\lambda}(y) : \hat{\beta}_{j,\lambda}(y) \neq 0 \}. \]  

(1.7)

To ease notation, we use the shorthands

\[ \hat{\beta}(y) = \hat{\beta}_{\lambda}(y) \]
\[ \hat{E} = \hat{E}_{\lambda}(y) \]
\[ \hat{z}_{E} = \hat{z}_{E,\lambda}(y). \]

(1.8)

In Section 2 we investigate the KKT (Karush-Kuhn-Tucker) conditions for the program (1.4). As in the LASSO case, the KKT conditions provide the basic description for the selection event on which selective inference is based.

1.2. Selective inference

Fithian et al. (2014) describe a broad framework for post-selection framework, most notably including exponential families such as the canonical regression model (1.1). This framework begins by specifying a set of statistical questions

\[ Q = \{(M_i, H_{0,i}), i \in I\}. \]  

(1.9)

Above, \( M \) is a statistical model (often parametric), and \( H_0 \subset M \) a null hypothesis concerning the model \( M \). A selection procedure can be formalized as a point process \( \hat{Q} \) taking values in \( Q \), determining which questions will be asked by an analyst.

The selective inference framework attempts to control the selective type I error rate (1.10). This is defined in terms of a pair of \( (M, H_0) \in Q \) and a critical function \( \phi_{(M,H_0)} \) to test \( H_0 \subset M \) vs. \( H_a = M \setminus H_0 \). Formally, the selective type I error is

\[ P_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected}) = E_{M,H_0} \left( \phi_{(M,H_0)}(y) \mid (M, H_0) \in \hat{Q}(y) \right). \]  

(1.10)

The process \( \hat{Q} \) determines a map taking a distribution \( P \in M \) to

\[ P \left( \cdot \mid (M, H_0) \in \hat{Q} \right). \]  

(1.11)

We call such distributions selective distributions, and define a selective model to be the collection of selective distributions

\[ \left\{ P \left( \cdot \mid (M, H_0) \in \hat{Q} \right) : P \in M \right\}. \]  

(1.12)

To make this concrete we now give examples in the regression setting. Model and hypothesis pairs \((M, H_0)\) are often of the form:

**Saturated model, \( \sigma_E^2 \) known:**

\[ M = M_s = \{ N(\mu, \sigma_E^2 I) : \mu \in \mathbb{R}^n \}, \quad H_0 = \{ \mu : \eta^T \mu = 0 \}. \]  

(1.13)

**Regression model, \( \sigma_E^2 \) known:**

\[ M = M_{k,E} = \{ N(X_E \beta_E, \sigma_E^2 I) : \beta_E \in \mathbb{R}^E \}, \quad H_0 = \{ \beta_E : \eta^T \beta_E = 0 \}. \]  

(1.14)
Regression model, $\sigma_E^2$ unknown:

$$M = M_{u,E} = \{N(X_E\beta_E, \sigma_E^2 I) : \beta_E \in \mathbb{R}^E, \sigma_E^2 \geq 0\}, \quad H_0 = \{\beta_E : \eta^T \beta_E = 0\}. \quad (1.15)$$

The goal of selective inference, as formulated above, is to build tests $\phi_{(M,H_0)}$ such that

$$\mathbb{E}_{M,H_0} \left( \phi_{(M,H_0)}(y) \mid (M,H_0) \in \hat{Q}(y) \right) \leq \alpha. \quad (1.16)$$

In the regression context with $\sigma_E^2$ known, the set of questions is most naturally chosen to be

$$Q_k = \{(M_k,E, \{\beta_E : \beta_j,E = 0\}) : E \subset \{1, \ldots, p\}, j \in E\}. \quad (1.17)$$

If $\sigma_E^2$ is unknown, we might choose

$$Q_u = \{(M_{u,E}, \{\beta_E : \beta_j,E = 0\}) : E \subset \{1, \ldots, p\}, j \in E\}. \quad (1.18)$$

With these choices of question spaces, the selection procedure $\hat{Q}$ is one that chooses a subset of variables to form a model (1.1). In terms of the square-root LASSO the most natural choice is perhaps:

$$\hat{Q}(y) = \hat{Q}_{u,\lambda}(y) = \{(M_{u,E}, \{\beta_E : \beta_j,E = 0\}) : j \in \hat{E}\}. \quad (1.19)$$

Alternatively, we might be interested in one-sided hypothesis tests in which case we might look consider

$$\hat{Q}(y) = \hat{Q}_{u,\lambda}(y) = \{(M_{u,E}, \{\beta_E : \text{sign}(\beta_j,E) \neq \hat{z}_{E,j}\}) : j \in \hat{E}\}. \quad (1.20)$$

One of the take-away messages from Fithian et al. (2014) is that there is a concrete procedure to follow for selective inference in (many) exponential families, particularly when the null hypotheses can be expressed in terms of a one-parameter subfamily of the natural parameter space of the exponential family. If $T(y)$ is the corresponding sufficient statistic, and $U(y)$ denotes the sufficient statistics corresponding to nuisance parameters, then the key quantity in such procedures is the law

$$\mathcal{L}_{(M,H_0)} \left( T(y) | U(y), (M,H_0) \in \hat{Q}(y) \right). \quad (1.21)$$

While our models above are not parameterized by the natural parameters, hypotheses regarding $\beta_E$ can still be easily carried out. In the regression context and questions generated as (1.19), the above laws are equivalent to

$$\mathcal{L}_{(M_{u,E},\beta_j,E = 0)} \left( \hat{\beta}_{j,E}^{OLS}(y) | X_{E,j}^T y, \| (I - P_{E \setminus j}) y \|_2^2, \hat{E}(y) = E \right) \quad (1.22)$$

where

$$\hat{\beta}_{j,E}^{OLS}(y) = X_{E,j}^T y. \quad (1.23)$$

are the ordinary least squares estimates of $\beta_E$ in the model $M_{u,E}$.

The questions space (1.20) is useful for one-sided tests and for computational reasons described in Lee et al. (2013). In this case we have the laws

$$\mathcal{L}_{(M_{u,E},\beta_j,E = 0)} \left( \hat{\beta}_{j,E}^{OLS}(y) | X_{E,j}^T y, \| (I - P_{E \setminus j}) y \|_2^2, (\hat{E}(y_1), \hat{z}_{E}(y_1)) = (E, z_E) \right). \quad (1.24)$$

These laws are described in detail in Section 3. We will see that they are truncated $T$ distributions with the degrees of freedom one would expect. Based on these laws, we construct exact tests for the coefficients $\beta_E$. Given that the appropriate laws in the case of $\sigma$ known are truncated Gaussian it is not surprising that the appropriate distributions here are truncated $T$ distributions. To construct selective intervals, we suggest a natural Gaussian approximation to the truncated $T$ distribution and investigate its performance in a regression problem.
1.3. Data carving and selectively unbiased estimation

In Fithian et al. (2014) it was noted that data carving—where some amount of data is held out for inference in the second stage—can often improve power at some small cost in model selection performance. In this setting, the appropriate conditional distributions are

\[ L(M_{u,E}, \beta, \sigma^2 = 0) \left( \beta^{OLS}_{E}(y) | X_{E,2}^T y, \|I - P_{E,2}\|_2^2, (\hat{E}(y_1), \tilde{z}_E(y_1)) = (E, z_E) \right), \tag{1.25} \]

where \((y_1, X_1)\) is a subset \(n_1 < n\) observations we use to select the model.

In Section 5 we describe the sampling algorithm used in Fithian et al. (2014) in more detail and consider data carving for the square-root LASSO using a similar Gaussian approximation to the case with no data heldout. As in the case of \(\sigma^2_k\) known, we observe an increase in power when holding out some data.

For our approximation we need some estimate of \(\sigma^2_k\). If enough data is held out in the first stage, there are in fact selectively unbiased estimates of the parameters \((\beta_E, \sigma^2_E)\). The following lemma implies the existence of selective UMVU estimators for these parameters. We state the lemma in the context of data carving and regression for clarity, though it clearly holds in more generality.

**Lemma 1.1** (Selective UMVU (W. Fithian, personal communication)). Suppose \(n_2 = n - n_1 < |E|\).

Then, for any distribution in \(P_E \in M_{u,E}\) the following are unbiased estimators

\[ \hat{\beta}_{E,2}(y) = X_{2,E}^T y \\
\hat{\sigma}^2_{E,2}(y) = \frac{\|I - P_{E,2}\|_2^2 y_2}{n_2 - |E|} \]

for \((\hat{\beta}_E, \hat{\sigma}^2_E)\) respectively.

Further, for any distribution \(P_E \in M_{u,E}\) define the selective distribution

\[ \hat{Q}_{E,z_E}(\cdot) = P_E(\cdot | (\hat{E}(y_1), \tilde{z}_E(y_1)) = (E, z_E)). \]

Then, \(\hat{\beta}_{E,2}\) and \(\hat{\sigma}^2_{E,2}(y)\) are also unbiased estimators of \((\beta_E, \sigma^2_E)\) respectively under \(\hat{Q}_{E,z_E}\). Finally,

\[ E_{\hat{Q}_{E,z_E}}(\hat{\beta}_{E,2}|X_{E,2}^T y, \|y\|_2^2), E_{\hat{Q}_{E,z_E}}(\hat{\sigma}^2_{E,2}|X_{E}^T y, \|y\|_2^2) \]

are the UMVU estimators of \((\beta_E, \sigma^2_E)\) under the selective model \(\{\hat{Q}_{E,z_E} : P_E \in M_{u,E}\}\).

**Proof.** The first claim is obvious. The second claim is the formal justification of data splitting Cox (1975). The third claim is just the Rao-Blackwell theorem applied to the distribution \(\hat{Q}_{E,z_E}\) which has sufficient statistics \(X_{E,2}^T y, \|y\|_2^2\). \(\square\)

**Remark 1.2.** We call the estimators \(\hat{\beta}_{E,2}, \hat{\sigma}^2_{E,2}\) selectively unbiased estimators as they are unbiased under the selective distribution \(\hat{Q}_{E,z_E}\). The more general form of the lemma states that if \(\hat{\theta}\) is selectively unbiased estimator under a selective model \(M\) for a parameter \(\theta : M \rightarrow \mathbb{R}^k\) and \(T\) is sufficient for \(\theta\) for the selective model \(M\), then the selective UMVU is

\[ E_Q(\hat{\theta}|T) \]

where \(Q \in M\) is arbitrary.

We use the above lemma in Section 5 to estimate \(\sigma^2_E\) before testing hypotheses or constructing confidence intervals.
1.4. Summary

The take-away message of this paper is that selective inference with $\sigma^2$ unknown is possible in the $n < p$ scenario using the square-root LASSO. A simple example illustrating this is given in Section 4.1. Holding out data in the first stage provides a similar increase in power in the second stage.

In Section 2 we describe the square-root LASSO in more detail. In particular, we describe the selection events

$$\{y : (\hat{E}(y), \hat{z}_E(y)) = (E, z_E)\}, \quad E \subset \{1, \ldots, p\}, z_E \in \{-1, 1\}^E.$$

Following this, in Section 3 we turn our attention to the main inferential tool, the law (1.24). In this section, we also interpret these laws as selective pseudo-likelihoods which leads us to natural debiasing procedures for the parameters of the selected model. In the square-root LASSO case, this view yields a natural estimate of the noise variance which we compare to some other existing estimates based on the square root LASSO. In Section 4.2, we combine the inferential tool with the multiscale method for changepoint detection where the underlying noise is generally unknown. In Section 5 we derive the sampling scheme for data carving in both the LASSO problem and the square-root LASSO problem.

2. The Square Root LASSO

Recall the convex program (1.4)

$$\hat{\beta}(y) = \hat{\beta}_\lambda(y) = \arg \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2 + \lambda \cdot \|\beta\|_1$$

as well as our shorthand for the selected variables and signs (1.6), (1.7).

The Karush-Kuhn-Tucker (KKT) conditions characterize the solution as follows: $(\hat{\beta}(y), \hat{z})$ is a solution of (2.1) if and only if

$$\frac{X^T(y - X\hat{\beta}(y))}{\|y - X\hat{\beta}(y)\|_2} = \lambda \cdot \hat{z}$$

$$\hat{z}_j \in \begin{cases} \text{sign}(\hat{\beta}_j(y)) & \text{if } j \in \hat{E}(y) \\ [-1, 1] & \text{if } j \notin \hat{E}(y). \end{cases}$$

We see that our choice of shorthand for $\hat{z}_E$ simply corresponds to the $\hat{E}(y)$ coordinates of the subgradient of the $\ell_1$ norm. We note that, at the cost of some notation, the $\ell_1$ penalty above can be replaced with a weighted $\ell_1$ norm which will be used in our change-point model in Section 4.2.

Our first observation, which we had not found in the literature on square-root LASSO, is the following.

**Lemma 2.1.** For every $(E, z_E)$, on the event $\{(\hat{E}(y), \hat{z}_E(y)) = (E, z_E)\}$ the solutions of the LASSO and square-root LASSO are related as

$$\hat{\beta}(y) = \hat{\beta}_\lambda(y) = \hat{\beta}_{\gamma(y)}(y)$$
where
\[ \hat{\gamma}(y) = \lambda \hat{\sigma}_E(y) \cdot \left( \frac{n - |E|}{1 - \lambda^2 \| (X_E^T)^\dagger z_E \|_2^2} \right)^{1/2} \tag{2.5} \]
and
\[ \hat{\sigma}_E^2(y) = \frac{\| (I - X_E X_E^T) y \|_2^2}{n - |E|} = \frac{\| (I - P_E) y \|_2^2}{n - |E|} \tag{2.6} \]
is the usual ordinary least squares estimate of \( \sigma^2_E \) in the model \( M_{u,E} \).

Proof. On the event in question, we can rewrite the KKT conditions using the fact \( X_{E} \hat{\beta} = X_{E} \hat{\beta}_E \) as
\[ X_E (y - X_E \hat{\beta}_E(y)) = c_E(y) \cdot \lambda \cdot z_E \] \( \tag{2.7} \)
\[ X_E^T (y - X_E \hat{\beta}_E(y)) = c_E(y) \cdot \lambda \cdot z_E \] \( \tag{2.8} \)
\[ \text{sign}(\hat{\beta}_E(y)) = z_E, \quad \| \hat{z}_E \|_\infty < 1 \] \( \tag{2.9} \)
with
\[ c_E(y) = \| y - X_E \hat{\beta}_E(y) \|_2. \]

Comparing (2.7) with the KKT conditions of LASSO in Lee et al. (2013), this indicates \( \hat{\gamma}(y) = \lambda c_E(y) \). Following a standard calculation involving the KKT conditions of the LASSO, we can write
\[ y - X_E \hat{\beta}_E(y) = (I - P_E) y + \lambda \cdot c_E(y) \cdot (X_E^T \hat{\beta}_E) \] \( \tag{2.10} \)
Strictly speaking, this calculation involves defining the quantity
\[ \hat{\beta}_E(y) = (X_E^T X_E)^{-1} (X_E^T y - \lambda \cdot c_E(y) \cdot z_E) \]
which is equal to \( E \) coordinates of the square-root LASSO solution on the event in question but not for all \( y \).

Computing the squared Euclidean norm of both sides yields
\[ c_E^2(y) = \frac{\| (I - P_E) y \|_2^2}{1 - \lambda^2 \| w_E \|_2^2} = \frac{\| (I - P_E) y \|_2^2}{1 - \lambda^2 \| (X_E^T)^\dagger z_E \|_2^2}. \]

\[ \Box \]

2.1. A first example

Before we move on to the general case, it is helpful to look at the characterization of the selection event in the case of orthogonal design matrix. Suppose the design matrix \( X \in \mathbb{R}^{n \times p} \) has orthogonal columns, \( \hat{\beta}_E \) and \( c_E \) takes a simple form,
\[ \hat{\beta}_E = X_E^T y - \lambda c_E z_E, \quad c_E = \hat{\sigma}_E \cdot \sqrt{\frac{n - |E|}{1 - \lambda^2 |E|}}. \]
The selection event
\[ \{ y : \text{sign}(\hat{\beta}_E(y)) = z_E \}, \]
is decoupled into \(|E|\) constraints. For each \(i \in E\),
\[ \frac{z_i x_i^T y}{\hat{\sigma}_E(y)} \geq \lambda \sqrt{\frac{n - |E|}{1 - \lambda^2 |E|}}. \tag{2.11} \]

The left-hand side of (2.11) is closely related to the inference on \(\beta_i\), and follows a T-distribution with \(n - |E|\) degrees of freedom. The constraint (2.11) is a constraint on the usual T-statistic on the selection event which implies one should use the truncated T-distribution to test whether or not \(\beta_i = 0\).

2.2. Characterization of the selection event

We now describe the selection event explicitly for general design matrices, which will be used for deriving the law (1.24). From the proof for Lemma 2.1, we see that the event
\[ \{ y : \text{sign}(\hat{\beta}_E(y)) = z_E \} \tag{2.13} \]
is equal to the event
\[ \{ y : \hat{\sigma}_E(y) \cdot \alpha_{i,E} - z_{i,E} \cdot U_{E,i}(y) \leq 0, i \in E \} \tag{2.14} \]
where
\[ U_{E,i}(y) = \frac{e_i^T X_E^\dagger y}{\|e_i^T X_E^\dagger\|_2} \]
and
\[ \alpha_{i,E} = \lambda \cdot z_{i,E} \cdot \|e_i^T X_E^\dagger\|_2 \cdot (1 - \lambda^2 \|X_E^\dagger z_E\|_2^2)^{-1/2} e_i^T (X_E^\dagger X_E)^{-1} z_E. \]

While the expression is a little involved, it is explicit and easily computable given \((X_E^\dagger X_E)^{-1}\).

Let us now consider the inactive inequalities. The event
\[ \{ y : \|\hat{z}_{-E}\|_\infty < 1 \} \tag{2.15} \]
is equal to the event
\[ \left\{ y : \left| \frac{X_i^T (I - P_E)y}{\lambda \cdot c_E(y)} + X_i^T (X_E^\dagger)^\dagger z_E \right| < 1, \quad i \in -E \right\}. \tag{2.16} \]

For each \(i \in -E\) these are equivalent the intersection of the inequalities
\[ \left( \frac{1 - \lambda^2 \|X_E^\dagger z_E\|_2^2}{\lambda^2} \right)^{1/2} X_i^T U_{-E}(y) < 1 - X_i^T (X_E^\dagger)^\dagger z_E \]
\[ \left( \frac{1 - \lambda^2 \|X_E^\dagger z_E\|_2^2}{\lambda^2} \right)^{1/2} X_i^T U_{-E}(y) > -1 - X_i^T (X_E^\dagger)^\dagger z_E. \tag{2.17} \]

where
\[ U_{-E}(y) = \frac{(I - P_E)y}{\|(I - P_E)y\|_2}. \tag{2.18} \]
3. The Law $Q_{E,z_E}$

Recall our question generating process (1.19). On the event $\{y : (\hat{E}(y), z_E(y)) = (E, z_E)\}$ the model we use for inference is $M_{u,E}$. That is, we are interested in the parametric family

$$y|X_E \sim N(X_E\beta_E, \sigma_E^2 I)$$

with $(\beta_E, \sigma_E^2)$ unknown. Call this parametric family $P_E$. Under (1.19) the conditional law (1.24) is a conditional distribution of

$$\hat{Q}_{E,z_E}(A) = P_E(A|\hat{E}(y), \hat{z}_E(y)) = (E, z_E)).$$

Define also, for each $u \in \text{row}(I - P_E)$

$$Q_{E,z_E,u-E}(A) = P_E(A|\hat{E}(y), \hat{z}_E(y)) = (E, z_E)), (3.1)$$

and we need only consider $Q_{E,z_E,u-E}(y)$. The proof of the following is straightforward.

**Lemma 3.1.** Under any $P_E \in M_{u,E}$ the random variables $\hat{\beta}_E(y), \hat{\sigma}_E^2(y), U_{-E}(y)$ are independent. Under $\hat{Q}_{E,z_E}$ the pair $\hat{\beta}_E(y), \hat{\sigma}_E^2(y)$ is independent of $U_{-E}(y)$. Further, under each of $P_E, Q_{E,z_E,u-E}(y)$ and $\hat{Q}_{E,z_E}$ the pair $\hat{\beta}_E(y), \hat{\sigma}_E^2(y)$ are sufficient for $(\beta_E, \sigma_E^2)$ and $U_{-E}(y)$ is ancillary. Finally, the law of $(\hat{\beta}_E(y), \hat{\sigma}_E^2(y))$ under $Q_{E,z_E}$ is equal to the law under $Q_{E,z_E,u-E}(y)$.

The distribution $Q_{E,z_E}$ is slightly simpler than $\hat{Q}_{E,z_E}$ though it contains all of the information about $(\beta_E, \sigma_E^2)$. For this reason, we use this distribution for inference. Formally, for every $P_E \in M_{u,E}$ and each $u \in E$, the law $Q_{E,z_E,u-E}$ is equivalent to the law

$$\mathcal{L} \left( \hat{\beta}_E(z), \hat{\sigma}_E^2(z) | (\hat{E}(y), \hat{z}_E(y)) = (E, z_E), U_{-E}(z) = U_{-E}(y) \right), \quad z \sim P_E.$$

In what follows, we condition on $U_{-E}(y)$ and write $Q_{E,z_E}$ dropping the $U_{-E}(y)$ in the notation.

### 3.1. Inference under quasi-affine constraints

The general form of the law $Q_{E,z_E}$ is that of a multivariate Gaussian and an independent $\chi^2$ of some degrees of freedom and satisfying some constraints. These constraints are affine in the Gaussian fixing the $\chi$, but not affine in the data. Specifically, suppose $y \sim N(\mu, \sigma_E^2 I)$ and write

$$C = -\text{diag}(z_E)X_E^T,$$

then

$$Cy \leq \hat{\sigma}_P(y) \cdot b$$

with

$$\hat{\sigma}_P^2(y) = \frac{\| (I - P)y \|^2_2}{\text{Tr}(I - P)}$$

and $CP = P, P\mu = \mu$ for some projection matrix $P$. Here, we have assumed $y \sim N(\mu, \sigma^2 I), P\mu = \mu$. As $(\mu, \sigma^2)$ vary, we see that this is simply a convenient reparametrization of the family $M_{u,E}$ assuming that $\text{rank}(X_E) = |E|$. Note that if $b \leq 0$, then the above constraints are convex in $y$ but otherwise the set is generally not convex.
Our goal is exact inference for $\eta^T \mu$, for $\eta$ satisfying $P\eta = \eta$, under
\[
P(y \in A | Cy \leq \bar{y}_P(y) \cdot b) = \mathcal{M}_{(C,b,P)}(A), \quad y \sim N(\mu, \sigma^2 I), P\mu = \mu.
\]
Without loss of generality we assume $\|\eta\|_2^2 = 1$.
In particular, we would like to test the null hypothesis $H_0 : \eta^T \mu = 0$ under $\mathcal{M}_{(C,b,P)}$. Following Fithian et al. (2014) and our discussion above, the relevant law is
\[
\eta^T y \big| (P - \eta\eta^T)y, \|y\|_2^2 \quad y \sim \mathcal{M}_{(C,b,P)}.
\] (3.3)
To test the null hypothesis $H_0 : \eta^T \mu = \theta$, the relevant law, after reparameterization is
\[
\eta^T y - \theta \big| (P - \eta\eta^T)(y - \theta), \|y - \theta\|_2^2 \quad y \sim \mathcal{M}_{(C,b,P)}.
\] (3.4)
Our main result in this section is that this conditional law is a truncated $T$ with $\text{Tr}(I - P)$ degrees of freedom with an explicitly computable truncation set.
For some set $\Omega \subset \mathbb{R}$ let $T_{\nu}|\Omega$ denote the distribution function of the law of $T_{\nu}|T_{\nu} \in \Omega$ :
\[
T_{\nu}|\Omega(t) = \mathbb{P}(T_{\nu} \leq t | T_{\nu} \in \Omega). \quad (3.5)
\]
**Theorem 3.2 (Truncated t).** Suppose that $y \sim \mathcal{M}_{(C,b,P)}$. The law
\[
\eta^T y - \theta \big| (P - \eta\eta^T)(y - \theta), \|y - \theta\|_2^2 \overset{D}{=} T_{\text{Tr}(I - P)|\Omega}
\] (3.6)
where
\[
\Omega = \Omega(C, b, P, \|(I - P)y\|_2^2 + (\eta^T y - \theta)^2, (P - \eta\eta^T)(y - \theta) , \theta).
\] (3.7)
The precise form of $\Omega$ is given in (3.8) below.
**Proof.** Following Fithian et al. (2014), we choose a convenient parameterization of the exponential family $\mathcal{M}_{(C,b,P)}$. We take these to be
\[
(U_{\theta}, V_{\theta}, W_{\theta})(y) = (\eta^T y - \theta, (P - \eta\eta^T)(y - \theta\eta), \|(I - (P - \eta\eta^T)(y - \theta\eta))\|_2^2)
\]
\[
= (\eta^T y - \theta, (P - \eta\eta^T)y, \|(I - (P - \eta\eta^T))y - \theta\eta\|_2^2).
\]
Note that
\[
W_{\theta}(y) = \|(I - P)y\|_2^2 + (\eta^T y - \theta)^2 = \|P^{1/2}y\|_2^2 + (\eta^T y - \theta)^2.
\]
In these parameters, we can rewrite the quasi-affine inequalities as
\[
U_{\theta}(y)\alpha + \xi \leq d^{-1/2}(W_{\theta}(y) - U_{\theta}(y))^1/2b
\]
with $d = \text{Tr}(I - P)$ the degrees of freedom and
\[
\alpha = C\eta, \quad \xi = \xi(V(y)) = C(\theta\eta + V(y)).
\]
In turn, these inequalities can be rewritten in terms of the usual $T$-statistic
\[
T_{\theta}(y) = \frac{\eta^T y - \theta}{\sigma_P(y)}.
\]
Specifically,

\[ T_\theta(y)W_\theta(y)^{1/2} \alpha + \xi \cdot \sqrt{d + T_\theta^2(y)} \leq W_\theta(y)^{1/2}b \]

with \( T_\theta \sim T_d \) independent of \( \xi(V(y)) \) and \( W_\theta(y) \sim \sigma^2 \cdot \chi_{d+1}^2 \left( \frac{(y^T(\mu - \theta))^2}{\sigma^2} \right) \) under the unconstrained law \( y \sim N(\mu, \sigma^2 I), P\mu = \mu \). Therefore, under \( M_{(C,b,P)} \) and assuming \( H_0 \) is true, the law of \( T_\theta \) is \( T_d \mid \Omega \) where

\[
\Omega(C,b,P,w,v,\theta) = \bigcap_{1 \leq i \leq \text{nrow}(A)} \left\{ t \in \mathbb{R} : t \sqrt{w \cdot \alpha_i + \xi_i(v) \cdot \sqrt{d + t^2}} \leq \sqrt{w \cdot b} \right\} \quad (3.8)
\]

and \( d, \xi(v) \) are as above. Each individual inequality can be solved explicitly, with each one yielding at most 2 intervals. In practice, we have observed the intersection of the above is not too complex. For convenience above we assumed that \( CP = P \). This happens to be true in our selection event (1.26). However, there may be other selection procedures for which this is not true. What this means is that our selection event has enforced some restrictions on the residual vector as well as the vector of fitted values. If we define

\[ U_-(y) = \frac{P_- y}{\|P_- y\|_2}. \]

Then, a straightforward calculation shows that the inequalities are equivalent to

\[ T_\theta(y)W_\theta(y)^{1/2} \alpha + \xi(V(y)) \cdot \sqrt{d + T_\theta^2(y)} \leq W_\theta(y)^{1/2}(b - d^{1/2}CU_- (y)). \]

One can use this representation of the inequalities defining the conditioning event in \( M_{(C,b,P)} \) to sample from the joint law of

\[ T_\theta(y), U_-(y) \mid V(y), W_\theta(y), \quad y \sim M_{(C,b,P)}. \]

to find the marginal distribution of \( T_\theta \).

3.2. Inference for \( \sigma \) and debiasing under \( M_{(C,b,P)} \)

The law \( M_{(C,b,P)} \) is parametric, and in the context of model selection we have observed \( y \) inside the set

\[ C y \leq \hat{\sigma}_P(y) \cdot b. \]

The usual OLS estimates \( X_{E}^iy \) are biased under \( M_{(C,b,P)} = M_{(C,b,P)}(\mu, \sigma^2) \). In this parametric setting, there is a natural procedure to attempt to debias these estimators.

If we fix the sufficient statistics to be \( T(y) = (Py, \|y\|^2_2) \), then the natural parameters of the laws in \( M_{(C,b,P)} \) are \( (\mu/\sigma^2, -(2\sigma^2)^{-1}) \). Solving the score equations

\[
\int_{\mathbb{R}^n} T(z) M_{(C,b,P)}(\mu, \sigma^2)(dz) - T(y) = 0 \quad (3.9)
\]

for \((\hat{\mu}(y), \hat{\sigma}^2(y))\) corresponds to (selective) maximum likelihood estimation under \( M_{(C,b,P)} \). In the orthogonal design and known variance setting, this problem was considered by Reid et al. (2013). In our current setting, this requires sampling from the constraint set, which is generally non-convex.
Instead, we consider estimation of each parameter separately based on a form of pseudo-likelihood. Unfortunately, in the unknown variance setting, this approach yields estimates either for coordinates of $\mu/\sigma^2$ or $\sigma^2$ rather than coordinates of $\mu$ itself. We propose estimating $\sigma^2$ using pseudo-likelihood and plugging in this value to a quantity analogous to $M_{(C,b,P)}$ but with known variance, i.e. the law of $y \sim N(\mu, \sigma^2 I)$, $P\mu = \mu$ with $\sigma^2$ known subject to an affine constraint. This approximation is discussed in Section 3.5 below.

The pseudolikelihood is based on the law of one sufficient statistic conditional on the other sufficient statistics. Therefore, to estimate $\sigma^2$ we consider the likelihood based on the law

$$\| (I - P)y \|^2_2 | Py, y \sim M_{(C,b,P);(\mu,\sigma^2)}.$$

(3.10)

This law depends only on $\sigma^2$ and can be used for exact inference about $\sigma^2$, though for the parameter $\sigma^2$ an estimate is perhaps more useful than selective tests or selective confidence intervals.

Direct inspection of the inequalities yield that this law is equivalent to $\sigma^2 \cdot \chi^2_{\Omega = (I - P)}$ truncated to the interval $[L(Py), U(Py)]$ where

$$L(Py) = \max_{i:b_i \geq 0} \frac{Cy_i}{b_i},$$

(3.11)

$$U(Py) = \min_{i:b_i < 0} \frac{Cy_i}{b_i}.$$

For $\Omega \subset \mathbb{R}$, let $G_{\nu,\sigma^2,\Omega}$ denote the law $\sigma^2 \cdot \chi^2_{\Omega}$ truncated to $\Omega$

$$G_{\nu,\sigma^2,\Omega}(t) = \mathbb{P} \left( \chi^2_{\nu} \leq t | \chi^2_{\nu} \in \sigma^2 \Omega \right).$$

The pseudolikelihood estimate $\hat{\sigma}_{PL}(y)$ for $\sigma^2$ is the root of

$$\sigma \mapsto H_{\chi^2_{\Omega = (I - P)}}(L(Py), U(Py), \sigma^2) - \hat{\sigma}_{PL}^2(y)$$

(3.12)

where

$$H_{\nu}(L, U, \sigma^2) = \frac{1}{\nu} \int_{[0, \infty]} t G_{\nu,\sigma^2,|L,U|}(dt).$$

This is easily solved by sampling from $\sigma^2 \chi^2_{\Omega = (I - P)}$ truncated to $[L(Py), U(Py)]$.

This procedure is illustrated in Figure 1a. Note that for observed values of $\hat{\sigma}_{PL}^2(y)$ near the truncation boundary the estimate varies quickly with $\hat{\sigma}_{PL}^2$ due to the plateau at the upper limit. We remedy this in two simple steps. First, we use a regularized estimate of $\sigma$ under this pseudolikelihood. Next, we apply a simple bias correction to this regularized estimate so that when $[L, U] = [0, \infty)$ we recover the usual OLS estimator. Specifically, for some $\theta$ we obtain a new estimator as the root of

$$\sigma \mapsto H_{\nu}(L, U, \sigma^2) + \theta \cdot \sigma^2 - (1 + \theta)\hat{\sigma}_{PL}^2(y)$$

We call this regularized pseudo-likelihood estimate $\hat{\sigma}_{PL,R}^2(y)$. In practice, we have set $\theta = \nu^{-1/2}$ so that this regularization becomes negligible as the degrees of freedom grows. The regularized estimate can essentially be thought of as the MAP from an improper prior on the natural parameter for $\sigma^2$. In this case, if $\delta = 1/(2\sigma^2)$ is the natural parameter the prior has density proportional to $\delta^{\nu - \theta}$. As $H_{\nu}(0, \infty, \sigma^2) = \sigma^2$, it is clear that in the untruncated case we recover the usual OLS estimator $\hat{\sigma}_P(y)$. 

3.3. Comparison of estimators

As the selection event for the square-root LASSO yields a law of the form $M_{(C,b,P)}$, we can study the accuracy of the estimator by comparing it to other estimators for $\sigma$ in the LASSO literature. We compared our estimator $\hat{\sigma}_{PL,R}$ to a few other estimates found in the literature:

- OLS estimator in the selected model, where $\sigma = \frac{\| (I-P_E)y \|}{n-|E|}$, $E$ is the active set.
- Scaled LASSO in Sun & Zhang (2011).
- Residual sum of squares based estimator using Lasso coefficients with $\lambda$ selected by cross validation Reid et al. (2013).

The designs were $200 \times 400$ generated from an equicorrelated Gaussian with correlation 0.3, columns normalized to have length 1. The sparsity was set to 20 non-zero coefficients each 7.5 in absolute value but with a random sign. The parameter $\kappa$ in the square-root LASSO was set to 0.5. With these settings, the square-root LASSO “screened”, or discovered a superset of the 20 non-zero coefficients with a success rate of approximately 56%.

The performance of the estimators was evaluated by considering the ratio $\hat{\sigma}(y)/E(\mu)$ where

$$E(\mu) = (\sigma^2 + \|(I-P_E)\mu\|_2^2)^{1/2}$$

is the square-root of the (non-selective) expected value of $\hat{\sigma}_E^2(y)$ under the true data generating distribution.

We see from Figure 2a that our estimator is approximately unbiased when we correctly recovered all variables, outperforming the other estimators. While in the case of partial recovery (non-screening), we see that the estimator is quite close to the expected value of the usual OLS estimate had we not chosen the model with the square-root LASSO. We include the estimator from Reid et al. (2013) only in the plot where the square-root LASSO screens, as it provides a comparison of the variability of the estimator. The pseudo-likelihood estimator seems less variable than Minimum CV estimator.
In fact, when \((y_i, X_i)\) are independent draws from a fixed Gaussian distribution, then for any \(E\), the model \(M_{u,E}\) is correctly specified in the sense that the law of \(y|X[, E]\) is contained in \(M_{u,E}\). In this setting the quantity \(E(\mu)\) is an asymptotically correct estimator \(\sigma^2_E = \text{Var}(y_i|x_i,E)\). Hence, we see that the pseudo-likelihood estimator may be considered a reasonable estimator when the square-root LASSO does not actually screen, i.e. find a superset of the “true” variables.

### 3.4. Regression diagnostics

Recall the scaled residual vector \(U_{E}(y)\) in (2.18) is ancillary under the laws \(P_E\) and \(Q_{E, E}\). \(U_{E}(y)\) follows a uniform distribution on the \(n\)-dimensional unit sphere intersecting the subspace determined by \(I - P_E\), truncated by the observed constraints (2.17). As \(U_{E}(y)\) is ancillary, we can sample from its distribution to carry out any regression diagnostics or goodness of fit tests. For a specific example, we might consider the observed maximum of the residuals \(\|U_{E}(y)\|_{\infty}\).

Another natural regression diagnostic might be to test whether individual or groups of variables not selected improve the fit. Specifically, suppose \(G\) is a subset of variables disjoint from \(E\). Then, the usual \(F\) statistic for including these variables in the model is measurable with respect to \(U_{E}\):

\[
F_{G|G\cup E}(y) = \frac{\| (P_{G\cup E} - P_E) y \|_2^2 / |G|}{\| (I - P_{G\cup E}) y \|_2^2 / (n - |G \cup E|)}
\]

\[
= \frac{\| (P_{G\cup E} - P_E) U_{E}(y) \|_2^2 / |G|}{\| (I - P_{G\cup E}) U_{E}(y) \|_2^2 / (n - |G \cup E|)}.
\]

Therefore, a selectively valid test of

\[
H_0 : \beta_{G|G\cup E} = 0
\]
can be constructed by sampling $U_{-E}$ under its null distribution and comparing the observed $F$ statistic to this reference distribution. Details of these diagnostics are a potential area of further work.

### 3.5. A Gaussian approximation to $\mathbb{M}_{(C,b,P)}$

Suppose we have access to a reasonable estimate of $\sigma$ such as the estimate described above. On the set $\{ y : \hat{\sigma}_P(y) = \sigma_P \}$ we see that $y$ satisfies the affine constraints

$$Cy \leq \sigma_P b.$$ 

A distribution $Q_{E,z_E}$ derived from some $P_E \in M_{a,E}$ is used for inference about the parameters $\beta_E$. Let $\eta$ be the normalized linear functional determining $\beta_{j, E}^{OL}$, then for any value of $\theta$, this distribution is restricted to the sphere of radius $\| y - \theta X_j \|$ intersect the affine space $\{ z : X_{E \setminus j}^T z = X_{E \setminus j}^T y \}$. Call this set $S(\| P_{E \setminus j}(y - \theta \eta) \|, X_{E \setminus j}^T y)$ The restriction of $Q_{E,z_E}$ to $S(\| P_{E \setminus j}(y - \theta \eta) \|, X_{E \setminus j}^T y)$ is of course just the law $P_E$ restricted to $S(\| P_{E \setminus j}(y - \theta \eta) \|, X_{E \setminus j}^T y)$ intersect the selection event. For $|E|$ not large relative to $n$ by the classical Poincaré’s limit Diaconis & Freedman (1987), the law of $\eta^T y - \theta$ under $P_E$ restricted to $S(\| P_{E \setminus j}(y - \theta \eta) \|, X_{E \setminus j}^T y)$ that is close to a Gaussian with variance $\| P_{E \setminus j}(y - \theta \eta) \|^2/(n - |E| + 1)$. Thus, we might approximate its distribution under $Q_{E,z_E}$ by a truncated Gaussian.

Next we determine the truncation interval. We have observed $y$ satisfying some quasi-affine constraints involving $\hat{\sigma}_P^2(y)$. If we were to replace $\hat{\sigma}_P^2(y)$ by $\| P_{E \setminus j}(y - \theta \eta) \|^2/(n - |E| + 1)$ then the selection event would be affine in $\eta^T y$ with the right hand side being approximately the variance of $\eta^T y$ under $P_E$ restricted to $S(\| P_{E \setminus j}(y - \theta \eta) \|, X_{E \setminus j}^T y)$.

**Remark 3.3 (Approximate distribution).** Suppose we are interested in testing the hypothesis $H_0 : \eta^T \mu = \theta$ in the family $\mathbb{M}_{(C,b,P)}$ for some $\eta \in \text{row}(C)$ given an estimate $\hat{\sigma}_P^2(y)$ of $\sigma^2$. We propose using the distribution

$$\mathcal{L}(\eta^T z - \theta | (P - \eta \eta^T)z, C z \leq \hat{\sigma}_P(y) b), \quad z \sim N(\mu, \hat{\sigma}_P^2(y) I).$$

We condition on $(P - \eta \eta^T) z$ as we have assumed $P_\mu = \mu$ in defining $\mathbb{M}_{(C,b,P)}$ and this is a sufficient statistic for the unknown parameter $(P - \eta \eta^T) \mu$.

To validate the approximation, we use a similar simulation scenario in Section 7 of Fithian et al. (2014). We generate rows of the design matrix $X_{150 \times 200}$ from an equicorrelated multivariate Gaussian distribution with pairwise correlation $\rho = 0.3$ between the variables. The columns are normalized to have length 1. The sparsity level is 10, with each non-zero coefficient having value 6. Results are shown in Table 1.

| Level | Coverage |
|-------|----------|
| 0.85  | 0.886    |
| 0.90  | 0.910    |
| 0.95  | 0.944    |
| 0.98  | 0.981    |

*Table 1: Coverage of confidence intervals using Gaussian approximation based on forming 10000 intervals.*
4. Examples

We now consider two examples, one in which \( n > p \) involving in-vitro resistance of different HIV sequences to various anti-retroviral drugs. Our second example considers a multiscale change point problem in which \( p \approx n^2 \).

4.1. In-vitro HIV resistance

| Mutation | Naive OLS | Selective |
|----------|-----------|-----------|
| P41L     | 2.82e-01  | 2.59e-01  |
| P62V     | 1.33e-01  | 3.04e-01  |
| P65R     | 1.82e-25  | 6.82e-14  |
| P67N     | 4.57e-07  | 3.90e-02  |
| P69I     | 1.92e-09  | 4.60e-02  |
| P75I     | 4.78e-01  | 2.27e-01  |
| P77L     | 2.81e-01  | 2.81e-01  |
| P83K     | 2.51e-03  | 2.17e-01  |
| P90I     | 6.09e-05  | 5.40e-01  |
| P115F    | 1.52e-02  | 5.41e-01  |
| P151M    | 8.74e-03  | 2.68e-01  |
| P181C    | 4.16e-05  | 5.96e-03  |
| P184V    | 0.00e+00  | 7.28e-14  |
| P215F    | 2.52e-05  | 6.67e-01  |
| P215Y    | 3.00e-06  | 1.65e-01  |
| P219R    | 6.33e-04  | 1.84e-01  |

Table 2: P-values for coefficients selected by square-root LASSO in the selected model.

In Rhee et al. (2006), 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. We apply the square-root LASSO to the protease inhibitor subset of their data and report p-values for the selected mutations for one of the drugs, Lamivudine (3TC). Results from this analysis are shown in Table 2 and Figure 3.

4.2. Multiscale changepoint estimation

We illustrate the use of the square-root LASSO in a setting where \( p = O(n^2) \) so that it is generally impossible to have a good estimate of \( \sigma \). The setting we are interested in is changepoint estimation. We are inspired by the multiscale changepoint detection algorithms Frick et al. (2014), Chan & Walther (2011) though we transform this testing problem to an estimation problem.

We assume the data comes from some underlying function \( f : [a, b] \to \mathbb{R} \) which is piecewise constant with \( k \) change-points (We do not require prior knowledge on \( k \)). Our goal is to

- Determine where the change-points are.
- Get an estimate of the noise level \( \sigma \).

More specifically, we formalize the problem as follows: We discretize the interval \([a, b]\) into \( n \) parts. \( y \in \mathbb{R}^n \) is the observation on the grid over the interval \([a, b]\). \( X \in \mathbb{R}^{n \times p} \) is the matrix whose
Fig 3: Intervals for coefficients selected by square-root LASSO. For comparison, we have included the “relaxed” square-root LASSO coefficients as the gray bars. The corresponding intervals have no coverage guarantees as opposed to the red intervals which have conditional coverage 95%. Note the intervals can be somewhat long. This is addressed in Section 5 in which some data is heldout before forming intervals.

column is the indicator for some interval $I$ contained in $[a, b]$. We assume:

$$y = X\beta^0 + \epsilon, \quad \epsilon \sim N(0, \Sigma).$$  \hspace{1cm} (4.1)

Let $\mu(\beta^0) = X\beta^0$, the $\mu(\beta^0)$ is the discretization of the underlying function $f(x)$. Moreover, if $\beta^0$ is sparse, then $\mu$ and hence $f$ only has a few change points.

Suppose we knew $\sigma$, we can recover $\beta^0$ and make inferences based on solving the following problem:

$$\min_{\beta} \frac{1}{2} \|y - X\beta\|^2 + \sum_k \sum_{|I|=k} \sqrt{k} \left( \lambda + \sigma \sqrt{2 \log n} \right) |\beta_I|.$$ \hspace{1cm} (4.2)

Note that (4.2) is a balance between the goodness of the fit and sparsity of the solution imposed by the $l_1$ penalization that resembles that in LASSO (Tibshirani 1996). The motivation for solving (4.2) is that the minimum $\lambda$ that yields a nonzero $\beta$ has the form

$$\lambda_{\text{max}} = T(y) = \max_k \max_{I: |I|=k} \frac{\sum_{i \in I} y_i}{\sqrt{k}} - \sigma \sqrt{2 \log n}.$$ \hspace{1cm} (4.3)

This is exactly the penalized scan statistic used in (Chan & Walther 2011) to detect whether there is a (single) changepoint in the data. This corresponds to testing the global null in the problem
If we fix $\lambda$, by solving (4.2), we get an active set $E$, and signs $z_E$ based on which selective inference is possible.

However, with no prior knowledge of the noise level $\sigma$ (which is usually the case in practice), we form a weighted square-root LASSO problem analogous to (4.2),

$$
\min_\beta \sqrt{n} \cdot \|y - X\beta\|_2 + \sum_k \sum_{|I|=k} \sqrt{k} \left( \lambda + \sqrt{2 \log \frac{n}{k}} \right) |\beta_I|.
$$

### 4.3. KKT condition for the square-root lasso problem

The KKT conditions for problem (4.4) is,

$$
\sqrt{n}X^T(y - X\hat{\beta}) = Dc_E \hat{z},
$$

where $\hat{z}$ is the subgradient, $c_E = \|y - X\hat{\beta}\|_2$, 

$$
D = \text{diag}\{d_1, d_2, \ldots, d_p\}, \quad d_I = \sqrt{k} \left( \lambda + \sqrt{2 \log \frac{n}{k}} \right), \quad |I| = k.
$$

Noticing the similarity to (2.2), it is easy to adapt the selective inference framework to the case of weighted $l_1$ norm. We will not repeat the details here.

We denote by $\lambda_{\text{max}}$ the minimum $\lambda$ such that the solution for (4.4) is nontrivial. Then

$$
\lambda_{\text{max}} = T(y) = \max_k \max_{I:|I|=k} \frac{|\sum_{i \in I} y_i|/\sqrt{k}}{\|y\|_2/\sqrt{n}} - \sqrt{2 \log \frac{n}{k}}.
$$

This will be an important quantity to the screening scheme we discuss below. The test statistic $T(y)$, which is the dual seminorm of our penalty, is what connects our proposed procedure to multiscale change point detection as in Chan & Walther (2011), Frick et al. (2014). In this work, the authors consider using $T$ as a test statistic. Under $H_0: \mu \equiv 0$ the distribution of $\lambda_{\text{max}}$ has the distribution considered in these earlier works. In using the square-root LASSO instead of a LASSO with a regularization parameter proportional to known $\sigma$, we also allow the variance to be unknown.

### 4.4. Screening

For many applications of multiscale change-point detection, the design matrix $X \in \mathbb{R}^{n \times p}$, has $p = O(n^2)$ columns (which can easily be millions) and columns are highly correlated. In this case, we need to perform some screening procedure described in Algorithm 1.

### 4.5. CGH data

Copy number changes or alterations is a form of genetic variation in the human genome. DNA copy number alterations (CNAs) have been linked to the development and progression of cancer and many diseases. Array CGH log2-based intensity ratios provide useful information about genome-wide CNAs. In the following simulation, we use the CGH data downloaded from Snijders et al. (2001) and use the data for cell line GM03576 for chromosomes 1 through 23. The measurements are the log2-based test/reference intensity ratios across the chromosomes. Figure 4 gives a fitted line for the Log2ratios as well as an estimator for $\sigma$. 

---

 imsart-generic ver. 2008/08/29 file: paper.tex date: May 1, 2015
Algorithm 1 Solve multiscale change-point detection with screening.

\[ \lambda_0 \leftarrow \alpha \lambda_{\text{max}}, \text{for some } \alpha < 1. \]

\[ \lambda = 2 E \left[ \max_k \max_I |I| | \frac{\sum_{i \in I} y_i}{\sqrt{|I|}} - \frac{1}{\sqrt{n}} \frac{\lambda}{\|E\|} \right], \text{ where } \epsilon \sim N(0, I). \]

For each interval \( I \), \( \text{score}(I) \leftarrow \frac{\sum_{i \in I} y_i}{\sqrt{|I|}} - \frac{1}{\sqrt{n}} \frac{\lambda}{\|y\|} \). \( A \leftarrow \text{which(score} > \lambda_0) \).

Solve (4.4) at \( \lambda \) for \( X_A \), check the KKT conditions (4.5) for the full matrix \( X \).

\textbf{while} there are predictors violating the KKT conditions \textbf{do}

Repeat screening using the residual, get \( A \).

Solve (4.4) at \( \lambda \) for \( X_A \), check the KKT conditions (4.5) for the full matrix \( X \), replacing \( y \) with the residuals.

---

5. Data carving

In Fithian et al. (2014) the authors compared selective inference using the LASSO to data splitting, in which \( n_1 < n \) cases are used to select a model and tests and confidence intervals are produced based on the remaining \( n_2 = n - n_1 \) cases. This example was used partly to demonstrate that data splitting is inadmissible and can be dominated by a procedure that uses all \( n \) data points in the second stage, i.e. for inference. This was referred to as data carving. Interestingly, a second curious phenomenon was noted. For \( n = 100 \), a moderate sized data set, even having \( n_2 = 1 \) yielded a noticeable improvement in power. We suspect that this is related to the fact that the distribution used for inference in the second stage is no longer a truncated distribution when \( n_2 > 0 \). Rather, it is the convolution of a univariate Gaussian with the projection of a multivariate truncated Gaussian. Formal justification of this statement seems like an interesting problem. In this section we consider data carving the square-root LASSO problem. We begin with a description of data carving the LASSO, which has some interesting features itself, and was not discussed in detail in Fithian et al.
5.1. Data carving the LASSO

When $n_1 > 0$, we partition the data into two groups $(y_1, X_1)$ and $(y_2, X_2)$. Fitting the LASSO with a fixed $\gamma$ to $(y_1, X_1)$ and observing the signs and identities of the selected variables yields two sets of affine constraints Lee et al. (2013):

$$\text{diag}(z_E) \left( X_{1,E}^\dagger y_1 - \gamma (X_{1,E}^T X_{1,E})^{-1} z_E \right) \geq 0$$

$$\left\| X_{1,-E}^T ((I_1 - P_{1,E}) y_1 + \gamma (X_{1,E}^T)^\dagger z_E) \right\|_{\infty} \leq \gamma$$ (5.1)

where $I_1 = I_{n_1 \times n_1}$ and $P_{1,E} = X_{1,E} X_{1,E}^\dagger$. Though these affine constraints can be stacked into a large matrix, they are naturally partitioned into two sets: constraints on $X_{1,E}^\dagger y_1$ and $(I_1 - P_{1,E}) y_1$ which are of course independent under the linear regression model with design matrix $X_{1,E}$ and known variance $\sigma_k^2$, which we might denote by $M_{k,E}$. We might write these two constraints as

$$\text{diag}(z_E) X_{1,E}^\dagger y_1 \leq b_E$$
$$A_E^\dagger (I_1 - P_{1,E}) y_1 \leq b_E^\perp$$ (5.2)

Following the setup described in Fithian et al. (2014) the appropriate distribution for inference about $\beta_j$ for some $j \in E$ in the model $M_{k,E}$ is the conditional law

$$(e_j^T X_{1,E}^\dagger y, \text{diag}(z_E) X_{1,E}^\dagger y_1 \leq b_E, A_E^\dagger (I_1 - P_{1,E}) y_1 \leq b_E^\perp, \ y \sim M_{k,E})$$ (5.3)

Our first result shows that this law is equivalent to a lower dimensional truncated Gaussian. The importance of this result is that it reduces each sampling problem from a truncated Gaussian on $\mathbb{R}^n$ to one on $\mathbb{R}^{\min(n_2,|E|)+1}$.

**Lemma 5.1.** Inference for parameters $\beta_{j,E}$ in model $M_{k,E}$ is equivalent to the linear projection of a multivariate Gaussian of dimension $\min(n_2,|E|)+1$ truncated by a set of $|E|$ inequalities. Specifically, define

$$L_{(E,z_E)} = \begin{cases} 
\text{diag}(z_E) X_{1,E}^\dagger S_1 \\ X_{1,E}^\dagger \\
\text{diag}(z_E) X_{1,E}^\dagger S_1 \\
S_2 
\end{cases} \begin{cases} 
n_2 \geq |E| \\
n_2 < |E|.
\end{cases}$$ (5.4)

Then, for each $j \in E$ inference for $\beta_{j,E}$ is based on a truncated conditional of $T_{(E,z_E)} = L_{(E,z_E)} y$ subject to affine constraints $T_{(E,z_E)}[|E|] \leq b_E$.

**Proof.** First, note that $X_{1,E}^\dagger y$ and $X_{1,E}^\dagger y_1$ are independent of $(I_1 - P_{1,E}) y_1$ and hence the conditional law is equivalent to

$$L \left( e_j^T X_{1,E}^\dagger y, X_{1,E}^\dagger, \text{diag}(z_E) X_{1,E}^\dagger y_1 \leq b_E \right), \ y \sim M_{k,E}.$$ (5.5)

As $(E, z_E)$ are fixed throughout we drop the $E$ in the notation and assume $z_E = 1$ so that $\text{diag}(z_E) = I_1$. Let $P_{-j} = X_{-j} X_{-j}^\dagger$ denote the projectors onto the column spaces of $X_{-j}$ and $n_j = P_{-j} X_j / \|P_{-j} X_j\|_2^2 = e_j^T X_{1,E}^\dagger$. We also drop the index $j$, writing $P_{-}$ and $\eta$.
It will be useful to also have the matrices $S_1, S_2$ so that
\[
\begin{pmatrix}
X_1 \\
0
\end{pmatrix} = S_1^T X_1, \quad \begin{pmatrix}
0 \\
X_2
\end{pmatrix} = S_2^T X_2.
\]

The selection event, having dropped the second set of constraints, is
\[
\left\{ y : X_1^T S_1 y \leq b \right\}.
\]

Above, we distinguished between two cases: $n_2 \geq |E|$ and $n_2 < |E|$. In either case, the estimators $\hat{\beta}_1, \hat{\beta}$ are linear functions of $L(E,x_E)y = Ly$ where $L$ is defined above in (5.4). Note that $\text{row}(L) = \min(n_2, |E|) + |E|$.

Under the above model, we are interested in the law
\[
\mathcal{L}(\eta^T y | X_1^T S_1 y \geq b, P_- y).
\]

where $\eta^T y, X_1^T S_1 y$ and $P_- y$ are linear functions of $Ly$. Setting $u = P_- y$, the above law is supported on
\[
\left\{ y : X_1^T S_1 y \geq b, P_- y = u \right\} = \left\{ y : X_1^T S_1 ((I - P_-) y + u) \geq b, P_- y = u \right\}
\]
a set of codimension $|E| - 1$. In terms of $Ly$ it follows then that the above law is also supported on a set of codimension $|E| - 1$. That is, it is supported on a set of dimension $\min(n_2, |E|) + 1$.

**5.2. Data carving the square root LASSO**

In the case of unknown $\sigma$, from (2.14) we can rewrite the active constraints for the square-root Lasso problem as
\[
\tilde{C}_1 X_1^T y_1 \leq \tilde{\sigma}_{1,E}(y_1) \hat{b}_1, \quad \tilde{\sigma}_{1,E}(y_1) = \frac{\| (I - P_{1,E}) y_1 \|_2}{n_1 - |E|}, \tag{5.6}
\]

Similar to an argument in Lemma 3.1, conditioning on the ancillary statistic $U_{-E,1}(y_1)$ means we can drop the inactive constraints and just use (5.6) as the description of the selection event if we are only interested in inference for the parameters $\beta_E$. In this case $\tilde{C}_1, \hat{b}_1$ now have fewer rows and we denote these by $C_1, b_1$. Following our earlier notation, we denote by the law
\[
y | C_1 X_1^T y_1 \leq \tilde{\sigma}_{1,E}(y_1) b_1.
\]

by $M_{C_1, b_1, S_1^T (I_1 - P_{1,E}) S_1}$ and by $M_{E_{C_1, b_1, S_1^T (I_1 - P_{1,E}) S_1} \subset M_{C_1, b_1, S_1^T (I_1 - P_{1,E}) S_1}$ those distributions in $M_{C_1, b_1, S_1^T (I_1 - P_{1,E}) S_1}$ that are derived from taking a distribution in $M_{E, E}$ and conditioning on the quasi-affine constraints corresponding $C_1, b_1, S_1^T (I_1 - P_{1,E}) S_1$ and the ancillary statistic $U_{-E,1}$.

In general, it is possible that $\hat{E}(y_1) \geq n_2 = n - n_1$. In this case, there is insufficient data remaining to derive a selectively unbiased estimate of $\sigma_E^2$ as described in Lemma 1.1. In this situation, we use the pseudo-likelihood estimate of $\sigma_E^2$ described in Section 3. Otherwise, we use Lemma 1.1 to produce a selectively unbiased estimate of $\sigma_E^2$. 
Lemma 5.2. Suppose $|E| < n_2$. For any distribution $Q_E$ in $M_{C_1,b_1,S_1^T(I_1-P_{E,1})S_1}$ the following holds
\[ E_{Q_E} \left( E_{Q_E} (\hat{\beta}_{E,2}(y) | X_E^T y, \|y\|^2_2) \right) = \beta_E \]
\[ E_{Q_E} \left( E_{Q_E} (\hat{\sigma}_{E,2}^2(y) | X_E^T y, \|y\|^2_2) \right) = \sigma_E^2. \]

In order to compute the estimate, we must sample from any distribution in $M_{C_1,b_1,S_1^T(I_1-P_{E,1})S_1}$ restricted to the set $S(y) = \{ z : X_E^T z = X_E^T y, \|z\|^2_2 = \|y\|^2_2 \}$. Sufficiency implies this distribution does not depend on which distribution we choose.

A natural Gibbs sampling scheme alternates between drawing

**Step 1:** $\hat{\beta}_1(z) - \hat{\beta}(z) | \hat{\sigma}_{E,1}(z)$ This is a draw from the projection of a random vector uniformly distributed on a sphere in row$(I - P_{E} - S_1^T (I_1 - P_{E,1})S_1) \subset$ row$(I - P_{E})$ with radius
\[ \| (I - P_{E}) y \|^2_2 - \hat{\sigma}_{E,1}^2(z)(n_1 - |E|) \]
subject to affine constraints. The projection is onto the rowspace of
\[ P_{\Delta,1} = \text{row}(X_1^T S_1 - X_1). \]

**Step 2:** $\hat{\sigma}_{E,1}^2(z) | \hat{\beta}_1(z) - \hat{\beta}(z)$ This is a draw from a Beta$(|E|, n - 2|E|)$ random variable truncated to an interval determined by $P_{\Delta,1}(z)$.

Rather than using this Gibbs sampling scheme we estimate $\hat{\sigma}_E^2$ then use the Gaussian approximation in Remark 3.3 using the same simulation scenario discussed after Remark 3.3. The results are shown in Figure 5. To estimate $\sigma_E$, we have several choices. We could just use the pseudo-likelihood estimate of $\sigma_E$ based on the data in stage 1. When $n_2 > s$ we have a selectively unbiased estimate.
of $\sigma_E$ based on $(y_2, X_2)$. In our simulations, when $n_2 > s$, rather than use the UMVU estimate of 1.1 we formed a pooled estimate of the pseudo-likelihood estimate and the unbiased estimate based on $(y_2, X_2)$. This is faster, though likely less accurate than using the UMVU.

5.3. 3TC data revisited

| Mutation | Data splitting | Data carving |
|----------|----------------|--------------|
| P41L     | 8.47e-01       | 5.48e-01     |
| P62V     | 2.14e-02       | 3.73e-01     |
| P65R     | 8.10e-06       | 4.57e-13     |
| P67N     | 1.49e-04       | 2.36e-08     |
| P69i     | 3.64e-02       | 4.53e-05     |
| P77L     | 2.28e-01       | 6.75e-01     |
| P83K     | 6.89e-01       | 2.80e-01     |
| P115F    | 4.47e-01       | 2.01e-01     |
| P151M    | 9.05e-02       | 2.80e-01     |
| P181C    | 8.11e-03       | 2.52e-03     |
| P184Y    | 8.00e-00       | 8.88e-16     |
| P215Y    | 1.75e-01       | 1.89e-02     |
| P219R    | 8.03e-02       | 6.54e-01     |

Table 3

P-values for coefficients selected by square-root LASSO in the selected model holding 10% out in the first stage.

To illustrate the effect of holding out some data, we reanalyze the 3TC data using only 90% of the data to choose the model. The results are illustrated in Figure 6 and Table 3. We computed the selective UMVU estimators for $(\beta_E, \sigma_E^2)$ described in Lemma 1.1. The sampling was carried out using a sequential Monte Carlo scheme described in Golchi & Campbell (2014). Code used for the analysis of the 3TC data can be found at https://gist.github.com/d2bd88beeea03117be16.

References

Belloni, A., Chernozhukov, V. & Wang, L. (2010), ‘Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming’, ArXiv e-prints.

Chan, H. P. & Walther, G. (2011), ‘Detection with the scan and the average likelihood ratio’, arXiv preprint arXiv:1107.4344.

Cox, D. (1975), ‘A note on data-splitting for the evaluation of significance levels’, Biometrika 62(2), 441–444.

Diaconis, P. & Freedman, D. (1987), ‘A dozen de finetti-style results in search of a theory’, Annales de l’Institut Henri Poincaré. Probabilités et Statistique 23(2, suppl.), 397–423.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), ‘Least angle regression’, Annals of Statistics 32(2), 407–499.

Fithian, W., Sun, D. & Taylor, J. (2014), ‘Optimal inference after model selection’, arXiv:1410.2597 [math, stat]. arXiv: 1410.2597.

URL: http://arxiv.org/abs/1410.2597

Frick, K., Munk, A. & Sieling, H. (2014), ‘Multiscale change point inference’, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(3), 495–580.

Golchi, S. & Campbell, D. A. (2014), ‘Sequentially Constrained Monte Carlo’, ArXiv e-prints.
Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2013), ‘Exact inference after model selection via the lasso’, *arXiv preprint arXiv:1311.6238*.

Lee, J. D. & Taylor, J. E. (2014), ‘Exact post model selection inference for marginal screening’, *arXiv:1402.5596 [cs, math, stat]*.

URL: [http://arxiv.org/abs/1402.5596](http://arxiv.org/abs/1402.5596)

Lockhart, R., Taylor, J. E., Tibshirani, R. J. & Tibshirani, R. (2013), ‘A significance test for the lasso’, *arXiv: 1301.7161*.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. & Yu, B. (2012), ‘A unified framework for high-dimensional analysis of MM-Estimators with decomposable regularizers’, *Statistical Science 27*(4), 538–557.

URL: [http://projecteuclid.org/euclid.ss/1356098555](http://projecteuclid.org/euclid.ss/1356098555)

Lockhart, R., Taylor, J. E., Tibshirani, R. J. & Tibshirani, R. (2013), ‘A Study of Error Variance Estimation in Lasso Regression’, *ArXiv e-prints*.

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L. & Shafer, R. W. (2006), ‘Genotypic predictors of human immunodeficiency virus type 1 drug resistance’, *Proceedings of the National Academy of Sciences 103*(46), 17355–17360.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K. et al. (2001), ‘Assembly of microarrays for genome-wide measurement of dna copy number’, *Nature genetics 29*(3), 263–264.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K. et al. (2001), ‘Assembly of microarrays for genome-wide measurement of dna copy number’, *Nature genetics 29*(3), 263–264.

Sun, T. & Zhang, C.-H. (2011), ‘Scaled Sparse Linear Regression’, *ArXiv e-prints*.

Taylor, J., Lockhart, R., Tibshirani, R. J. & Tibshirani, R. (2014), ‘Post-selection adaptive inference for least angle regression and the lasso’, *arXiv:1401.3889 [stat]*.

URL: [http://arxiv.org/abs/1401.3889](http://arxiv.org/abs/1401.3889)

Taylor, J., Loftus, J., Tibshirani, R. & Tibshirani, R. (2013), ‘Tests in adaptive regression via the kac-rice formula’, *arXiv preprint arXiv:1308.3020*.

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

Tibshirani, R. J. (2013), ‘The lasso problem and uniqueness’, *Electronic Journal of Statistics 7*, 1456–1490.
Fig 6: Intervals for coefficients selected by square-root LASSO with 10% heldout at random. As point estimates, we have included OLS estimates from the second stage in data splitting as well as the UMVU estimate of $\beta$ described in Lemma 1.1. The data splitting intervals have the same coverage guarantees as the data carving intervals. However, generally speaking, the data carving intervals are shorter than the data splitting intervals as the data carving intervals uses all of the data for inference in the second stage. This is analogous to the increase noted in power in Figure 5.