Biometrika (2018), **00**, 0, *pp.* 1–14
Printed in Great Britain
doi: 10.1093/biomet/asy045

Selective inference with unknown variance via the square-root lasso

BY XIAOYING TIAN

Farallon Capital Management LLC, One Maritime Plaza, 21st Floor, San Francisco, California 94115, U.S.A.

xtian@faralloncapital.com

JOSHUA R. LOFTUS

Department of Information, Operations, and Management Sciences, New York University, 44 West Fourth Street, New York, New York 10012, U.S.A.

loftus@nyu.edu

AND JONATHAN E. TAYLOR

Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, California 94305, U.S.A.

jonathan.taylor@stanford.edu

SUMMARY

There has been much recent work on inference after model selection in situations where the noise level is known. However, the error variance is rarely known in practice and its estimation is difficult in high-dimensional settings. In this work we propose using the square-root lasso, also known as the scaled lasso, to perform inference for selected coefficients and the noise level simultaneously. The square-root lasso has the property that the choice of a reasonable tuning parameter does not depend on the noise level in the data. We provide valid p-values and confidence intervals for coefficients after variable selection and estimates for the model-specific variance. Our estimators perform better in simulations than other estimators of the noise variance. These results make inference after model selection significantly more applicable.

Some key words: Confidence interval; Hypothesis testing; Lasso; Model selection; Square-root lasso.

1. INTRODUCTION

1·1. *Inference after model selection*

We consider inference after regression model selection in situations where the error variance is unknown. Given a vector of outcomes $y \in \mathbb{R}^n$ and a design matrix of predictors $X \in \mathbb{R}^{n \times p}$, we first choose a model by selecting some subset S of the columns of X. Denoting this model submatrix by X_S, we proceed with the regression model

$$y = X_S \beta_S + \epsilon, \quad \epsilon \sim N(0, \sigma_S^2 I),$$

and conduct the usual inferential tasks undertaken in regression, such as hypothesis tests and the construction of confidence intervals. In this work, we consider the design matrix X to be fixed.

Inference is carried out using the framework of selective inference (Fithian et al., 2017; Lee et al., 2016), which differs from classical inference in two respects. First, the model is selected after initial inspection of the data. In the above formulation, we propose a model indexed by S after the data have been found to suggest that S are the variables worthy of further investigation. Second, inference is carried out conditionally on the event that columns in S are selected. Unlike classical inference, selective inference recognizes and adjusts for data exploration used to inform the analyst's choice of model.

We first review the general framework of selective inference. Suppose that our data D lie in some measurable space $(\mathcal{D}, \mathcal{F})$, with unknown sampling distribution $\mathbb{F}$. Selective inference allows the data analyst to choose a reasonable probability model M, i.e., a subset of the probability measures on $(\mathcal{D}, \mathcal{F})$, in a data-driven fashion, before carrying out inference in M. Type I error and coverage guarantees are said to hold selectively (Fithian et al., 2017; Lee et al., 2016). In such an approach, the data analyst must specify a variable selection algorithm, represented here by a set-valued map $\hat{\mathcal{Q}} : \mathcal{D} \to \mathcal{Q}$ where $\mathcal{Q}$ is a collection of statistical models.

In this work we have a fixed design matrix X, and so the data y lie in $\mathcal{D} = \mathbb{R}^n$. The variable selection algorithm is based on the square-root lasso of Belloni et al. (2011), also known as the scaled lasso of Sun & Zhang (2012). We discuss the details of the algorithm in § 1·2. Using the square-root lasso to select columns allows us to perform inference after selection even when the noise level is unknown. After selecting a subset of columns S using the square-root lasso, we consider the model indexed by S with Gaussian errors

$$M = M_S = \{N(X_S\beta_S, \sigma_S^2 I) : \beta_S \in \mathbb{R}^{|S|}, \sigma_S^2 \geqslant 0\}, \tag{1}$$

where $N(\cdot, \cdot)$ denotes the Gaussian distribution with associated mean and covariance parameters. Without loss of generality, we assume that $\mathcal{Q}$ is the collection of M_S for all subsets of the columns.

Model (1) is different from most of those proposed in the literature (Lee et al., 2016; Taylor et al., 2016; Tibshirani et al., 2016), as previous works assume that σ_S^2 is known. This assumption is problematic for two reasons. First, it is almost never true in practice. Second, the noise level σ_S as posited above is specific to the model we choose; namely, σ_S^2 is the sum of the irreducible noises and signals not captured by our sparse estimates indexed by S. As we choose the variables S using the data, it is generally not easy to get an independent estimate of σ_S. Here we propose a method that treats σ_S as one of the parameters for inference and adjusts for selection. Unknown variance is discussed and allowed for in Fithian et al. (2017), although theoretically optimal choices of the tuning parameter will also typically depend on this unknown quantity. The square-root lasso avoids this problem by having a scale-free choice of tuning parameter.

In general, there is no guarantee that M_S is the correct model. The variable selection procedure might not screen or include all the nonzero coefficients. Moreover, the normality assumptions can be violated. However, we shall show in § 4·2 that our method developed under Gaussian assumptions is robust with respect to model misspecification, even when both the normality and the screening assumptions are violated. In either case, we obtain a good estimator for σ_S^2, the model-specific variance.

With the conditional approach taken in selective inference, the argument is that as we have used the data to choose M_S, it is reasonable to consider distributions for inference conditioning on selection of the same model. We denote this law by

$$\mathcal{L}\{y \mid M_S \in \hat{\mathcal{Q}}(y)\}. \tag{2}$$

This law can be used to obtain inferences for both the coefficients β_S and the variance σ_S^2. To test a hypothesis concerning a single coefficient $\beta_{j,S}$, for any $j \in S$, or concerning the variance

parameter σ_S^2, we can further condition on the sufficient statistics for the nuisance parameters following the properties of exponential families. For details, see § 1·3.

Our method is valid in both theory and practice. We illustrate the latter through comparisons of estimates of σ_S^2 with those of Sun & Zhang (2012) and Reid et al. (2016), and comparisons of false discovery rate control and power with those of Barber & Candès (2018).

1·2. *Model selection with the square-root lasso*

The variable selection procedure we use is based on the square-root lasso of Belloni et al. (2011), which is equivalent to the scaled lasso of Sun & Zhang (2012). It is defined by the following optimization problem with a tuning parameter λ:

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \|y - X\beta\|_2 + \lambda \|\beta\|_1. \tag{3}$$

This is a modification of the lasso (Tibshirani, 1996),

$$\tilde{\beta}_\gamma = \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \frac{1}{2}\|y - X\beta\|_2^2 + \gamma \|\beta\|_1,$$

where γ is also a tuning parameter. We will see in § 2·1 that the lasso and square-root lasso have equivalent solution paths, and we derive an explicit map from λ to γ under which corresponding solutions to the optimization problems are identical.

A key advantage of using the square-root lasso is the convenience in choosing the penalty parameter λ. For the lasso, a good choice of γ depends on the noise variance σ_S. For example, Negahban et al. (2012) proposed

$$\gamma = 2E(\|X^{\mathrm{T}}\epsilon\|_\infty), \quad \epsilon \sim N(0, \sigma_S^2 I).$$

However, when σ_S is unknown, this formula is of little practical use. Good estimates of σ_S are difficult to obtain until we have chosen a good tuning parameter γ. This dilemma is resolved by the square-root lasso. Unlike with the lasso, the choice of tuning parameter λ for the square-root lasso does not depend on σ_S:

$$\lambda = \kappa E\left(\frac{\|X^{\mathrm{T}}\epsilon\|_\infty}{\|\epsilon\|_2}\right), \quad \epsilon \sim N(0, I) \tag{4}$$

for some constant κ. Below we use $\kappa \approx 1$. Typical theoretical guarantees require a value greater than 1, but in our simulations we have found that a choice slightly less than 1 also works quite well. Any spherically symmetric distribution yields the same choice of λ, so (4) is independent of σ_S.

Both the lasso and the square-root lasso can be viewed as model selection procedures $\hat{\mathcal{Q}}$, as they select variables to be included in the linear model. Following Tibshirani (2013), we assume that the columns of X are in general position to ensure uniqueness of solutions. This means that any affine subspace of $\mathbb{R}^n$ of dimension $k < n$ contains at most $k + 1$ columns from X or $-X$.

We define the selected variables for the square-root lasso as

$$\hat{S}_\lambda(y) = \{j : \hat{\beta}_{j,\lambda}(y) \neq 0\} \tag{5}$$

and the selected signs as

$$\hat{z}_{S,\lambda}(y) = \operatorname{sign}\{\hat{\beta}_{j,\lambda}(y) : \hat{\beta}_{j,\lambda}(y) \neq 0\}. \tag{6}$$

To simplify the notation, we write

$$\hat{\beta}(y) = \hat{\beta}_\lambda(y), \quad \hat{S} = \hat{S}_\lambda(y), \quad \hat{z}_{\hat{S}} = \hat{z}_{S,\lambda}(y).$$

Our selected model M_S is uniquely determined by the selected columns S, but we also condition on the signs to make computations easier, a practice also adopted in Lee et al. (2016) and Tibshirani et al. (2016).

In § 2 we investigate the Karush–Kuhn–Tucker conditions for the optimization problem in (3). These conditions provide the basic description of the selection event on which selective inference is based.

1·3. *Exponential family and its sufficient statistics*

Fithian et al. (2017) observed that if the underlying distribution of y belongs to an exponential family, then the conditional law in (2) also belongs to an exponential family. Moreover, the two exponential families share the same sufficient statistics and natural parameters and differ only in the reference measures. In this paper, our model M_S is a multi-parameter exponential family with natural parameters, up to scaling of sufficient statistics, given by

$$\left(\frac{\beta_S}{\sigma_S^2}, \frac{1}{\sigma_S^2}\right).$$

Hence, to perform inference for a single parameter, we can condition on the sufficient statistics corresponding to the nuisance parameters.

LEMMA 1. *For the regression model M_S in (1), we can test the hypothesis $H_0 : \beta_{j,S}/\sigma_S^2 = \theta$, for any $j \in S$, by considering the law*

$$\mathcal{L}_{(M_S, H_0)}\{X_j^{\mathrm{T}} y \mid \|y\|^2, X_{S\setminus j}^{\mathrm{T}} y, M_S \in \hat{\mathcal{Q}}(y)\}. \tag{7}$$

Similarly, the following law can be used for inference on σ_S^2:

$$\mathcal{L}_{(M_S, \sigma_S^2)}\{\|y\|^2 \mid X_S^{\mathrm{T}} y, M_S \in \hat{\mathcal{Q}}(y)\}. \tag{8}$$

By studying the distributions (7) and (8), we will be able to perform inference after selection via the square-root lasso. To gain insight into (7), we first look at the case without selection, where the law in (7) is that of the t-statistic

$$\frac{e_j^{\mathrm{T}} X_S^{\dagger} y}{\hat{\sigma}_S \|e_j^{\mathrm{T}} X_S^{\dagger}\|_2},$$

with degrees of freedom $n - |S|$, where

$$\hat{\sigma}_S^2 = \frac{\|(I - P_S)y\|^2}{n - |S|}, \quad X_S^{\dagger} = (X_S^{\mathrm{T}} X_S)^{-1} X_S^{\mathrm{T}}, \quad P_S = X_S X_S^{\dagger}.$$

Thus $X_S^{\dagger}$ is the pseudo-inverse of matrix X_S and P_S is the projection matrix onto the column space of X_S.

The selection event is equivalent to $\{\hat{S}(y) = S\}$, where $\hat{S}$ is defined in (5). We can explicitly describe the selection procedure if we further condition on the signs z_S; that is, we condition on the event $[\{\hat{S}(y), \hat{z}_{\hat{S}}(y)\} = (S, z_S)]$ in (7) and (8). Procedures valid under such laws are also valid under those conditioned on $\{\hat{S}(y) = S\}$, since we can always marginalize over $\hat{z}_{\hat{S}}$. As mentioned previously, for computational reasons we always condition on z_S as well.

In § 3 we will see that the inferential distributions are truncated t and χ^2 distributions with degrees of freedom $n - |S|$. Based on these laws, we construct exact tests for β_S and estimates for

σ_S. Given that the appropriate laws in the case of known σ are truncated Gaussian distributions, it is not surprising that the appropriate distributions here are truncated t distributions. To construct selective intervals, we suggest a natural truncated Gaussian approximation to the truncated t distribution and investigate its performance in simulations.

2. The square-root lasso

2·1. *Karush–Kuhn–Tucker conditions*

Recall the convex program (3) and our shorthand for the selected variables and signs in (5) and (6). The Karush–Kuhn–Tucker conditions characterize the solution as follows: $\{\hat\beta(y), \hat z\}$ is the solution and corresponding subgradient of (3) if and only if

$$\frac{X^{\mathrm{T}}\{y - X\hat\beta(y)\}}{\|y - X\hat\beta(y)\|_2} = \lambda \hat z, \tag{9}$$

$$\hat z_j \in \begin{cases} \operatorname{sign}\{\hat\beta_j(y)\}, & j \in \hat S(y), \\ [-1, 1], & j \notin \hat S(y). \end{cases} \tag{10}$$

Our shorthand for $\hat z_{\hat S}$ corresponds to the $\hat S(y)$ coordinates of the subgradient of the ℓ_1-norm.

Our first observation, which as far as we know has not appeared in the literature on the square-root lasso, is that the square-root lasso and the lasso, which have equivalent solution paths, share an explicit reparameterization of the solution paths.

Lemma 2. *For every (S, z_S), on the event $[\{\hat S(y), \hat z_{\hat S}(y)\} = (S, z_S)]$ the solutions of the lasso and the square-root lasso are related by $\hat\beta(y) = \hat\beta_\lambda(y) = \tilde\beta_{\hat\gamma(y)}(y)$, where*

$$\hat\gamma(y) = \lambda \hat\sigma_S(y) \left\{ \frac{n - |S|}{1 - \lambda^2 \|(X_S^{\mathrm{T}})^\dagger z_S\|_2^2} \right\}^{1/2},$$

$$\hat\sigma_S^2(y) = \frac{\|(I - X_S X_S^\dagger) y\|_2^2}{n - |S|} = \frac{\|(I - P_S) y\|_2^2}{n - |S|}.$$

Note that $\hat\sigma_S^2$ is the usual ordinary least squares estimate of σ_S^2 in the model M_S.

The proof of the lemma, which we defer to the Supplementary Material, is a direct result of the equality in (9) and its counterpart in the Karush–Kuhn–Tucker conditions for the lasso.

2·2. *Orthogonal case*

We now use the Karush–Kuhn–Tucker conditions to describe the selection event. Before we move on to the general case, it is helpful to look at the characterization of the selection event in the case of an orthogonal design matrix. When the design matrix $X \in \mathbb{R}^{n \times p}$ has orthogonal columns, $\hat\beta_S$ can be simplified to $\hat\beta_S = X_S^{\mathrm{T}} y - \lambda c_S z_S$ where

$$c_S = \|y - X\hat\beta(y)\|_2 = \hat\sigma_S \left(\frac{n - |S|}{1 - \lambda^2 |S|} \right)^{1/2}.$$

Because of the orthogonal design matrix X, the inactive block in (10) is independent of the active block. Thus the selection event is reduced to

$$\left[y : \operatorname{sign}\{\hat\beta_S(y)\} = z_S \right]$$

and can be decoupled into $|S|$ constraints given by, for each $i \in S$,

$$\frac{z_i x_i^{\mathrm{T}} y}{\hat{\sigma}_S(y)} \geqslant \lambda \left(\frac{n - |S|}{1 - \lambda^2 |S|} \right)^{1/2}. \tag{11}$$

The left-hand side of (11) is closely related to the inference on β_i, and follows a t distribution with $n - |S|$ degrees of freedom. Expression (11) constrains the usual t-statistic to the selection event, so one should use the truncated t distributions for inference on β_S.

2·3. *Characterization of the selection event*

We now characterize the selection event for general design matrices. This characterization will be used in deriving the laws (7) and (8). The proof of this lemma follows from the Karush–Kuhn–Tucker conditions. The active and inactive constraints correspond to the active and inactive blocks in (10).

LEMMA 3. *The variable selection procedure defined by the square-root lasso can be characterized as the intersection of the active constraints*

$$\{y : \hat{\sigma}_S(y)\alpha_{i,S} - z_{i,S} U_{i,S}(y) \leqslant 0,\ i \in S\}$$

and, writing $w_S = (X_S^{\mathrm{T}})^{\dagger} z_S$, *the inactive constraints*

$$\left\{ y : -1 - X_i^{\mathrm{T}} w_S < \left(\frac{1 - \lambda^2 \|w_S\|_2^2}{\lambda^2} \right)^{1/2} X_i^{\mathrm{T}} U_{-S}(y) < 1 - X_i^{\mathrm{T}} w_S,\ i \notin S \right\}, \tag{12}$$

where

$$U_{i,S}(y) = \frac{e_i^{\mathrm{T}} X_S^{\dagger} y}{\|e_i^{\mathrm{T}} X_S^{\dagger}\|_2}, \quad \alpha_{i,S} = \lambda z_{i,S} \left(\frac{n - |S|}{1 - \lambda^2 \|w_S\|_2^2} \right)^{1/2} \left\{ \frac{e_i^{\mathrm{T}} (X_S^{\mathrm{T}} X_S)^{-1} z_S}{\|e_i^{\mathrm{T}} X_S^{\dagger}\|_2} \right\}, \quad i \in S,$$

and

$$U_{-S}(y) = \frac{(I - P_S) y}{\|(I - P_S) y\|_2}. \tag{13}$$

Although the expression for $\alpha_{i,S}$ looks complicated, it is easily computed given $(X_S^{\mathrm{T}} X_S)^{-1}$, which need only be calculated once.

We can now simplify the laws (7) and (8) with the following lemma, which removes the inactive constraints (12).

LEMMA 4. *Conditional on* $\{\hat{S}(y), \hat{z}_{\hat{S}}(y) = (S, z_S)\}$, *for any* $m_S \in M_S$ *the law for inference of* $(\beta_S/\sigma_S^2, 1/\sigma_S^2)$ *is equivalent to*

$$\mathbb{Q}_{S,z_S} = \mathcal{L}_{m_S}\{U_S(y), \|(I - P_S) y\|^2 \mid \hat{\sigma}_S \alpha_{i,S} - z_{i,S} U_{i,S} \leqslant 0,\ i \in S\}. \tag{14}$$

The laws (7) *and* (8) *simplify to*

$$\mathcal{L}\{U_{S,j}(y) \mid \|y\|^2, U_{S,S\setminus j}(y), \hat{\sigma}_S \alpha_{i,S} - z_{i,S} U_{i,S} \leqslant 0,\ i \in S\},$$
$$\mathcal{L}\{\hat{\sigma}_S^2(y) \mid U_S(y), \hat{\sigma}_S \alpha_{i,S} - z_{i,S} U_{i,S} \leqslant 0,\ i \in S\}.$$

3. Inference under the conditional law

3·1. *Inference for β_S under quasi-affine constraints*

The law $\mathbb{Q}_{S,z_S}$ in (14) is that of independent multivariate Gaussian and $\chi^2_{n-|S|}$ variables subject to constraints. This motivates us to study the distribution of $y \sim N(\mu, \sigma^2 I)$ subject to the quasi-affine constraints

$$Cy \leqslant \hat{\sigma}_P(y) b \tag{15}$$

with

$$\hat{\sigma}^2_P(y) = \frac{\|(I-P)y\|_2^2}{\mathrm{tr}(I-P)},$$

where P is a projection matrix, $C \in \mathbb{R}^{q \times n}$, $b \in \mathbb{R}^q$ and

$$CP = C, \quad P\mu = \mu. \tag{16}$$

The assumptions in (16) are made to simplify notation. Inference for quasi-affine constraints without these assumptions can be made similarly. In inference after the square-root lasso, (16) is automatically satisfied. We write the law of y subject to these constraints as

$$\mathbb{M}_{C,b,P}(A) = \mathrm{pr}\{y \in A \mid Cy \leqslant \hat{\sigma}_P(y) b\}, \quad y \sim N(\mu, \sigma^2 I).$$

Our goal is exact inference for $\eta^{\mathrm{T}}\mu$, for a vector η satisfying $P\eta = \eta$. Without loss of generality we let $\|\eta\|_2^2 = 1$. To test the null hypothesis $H_0 : \eta^{\mathrm{T}}\mu = \theta$, we parameterize the data into $\eta^{\mathrm{T}}y$, the projection onto the direction η, and the orthogonal set $(P - \eta\eta^{\mathrm{T}})y$. With (16) and some algebra, we can see that the statistic $\eta^{\mathrm{T}}y$ is sufficient for $\eta^{\mathrm{T}}\mu$ and that $\{(P - \eta\eta^{\mathrm{T}})y, \|y\|^2\}$ is sufficient for the nuisance parameters. In the following, we will prove that the law of

$$\eta^{\mathrm{T}}y - \theta \;\mid\; (P - \eta\eta^{\mathrm{T}})y, \|y - \theta\eta\|_2^2, \quad y \sim \mathbb{M}_{C,b,P}$$

is truncated t with degrees of freedom $\mathrm{tr}(I-P)$ and an explicit truncation set.

For any set $\Omega \subset \mathbb{R}$ let $t_{d|\Omega}$ denote the distribution function of a truncated t distribution with d degrees of freedom and truncation region Ω.

Theorem 1. *Suppose that $y \sim \mathbb{M}_{C,b,P}$. The law of*

$$\eta^{\mathrm{T}}y - \theta \;\mid\; (P - \eta\eta^{\mathrm{T}})(y - \theta\eta), \|y - \theta\eta\|_2^2$$

is $t_{d|\Omega}$, where $d = \mathrm{tr}(I-P)$, and the form of the truncation set

$$\Omega = \Omega\{C, b, P, \|(I-P)y\|_2^2 + (\eta^{\mathrm{T}}y - \theta)^2, (P - \eta\eta^{\mathrm{T}})y, \theta\}$$

is given in (17) *below.*

Proof. Define notation for the sufficient statistics as follows:

$$(U_\theta, V, W_\theta)(y) = \big[\eta^{\mathrm{T}}y - \theta,\ (P - \eta\eta^{\mathrm{T}})y,\ \|\{I - (P - \eta^{\mathrm{T}}\eta)\}(y - \theta\eta)\|_2^2\big].$$

Since $W_\theta(y) = \|y - \theta\eta\|^2 - \|(P - \eta\eta^{\mathrm{T}})y\|^2$, conditioning on $(V, \|y - \theta\eta\|^2)$ is equivalent to conditioning on (V, W_θ).

Our main strategy is to construct a test statistic independent of (V, W_θ). This can be done through the usual t-statistic

$$\tau_\theta(y) = \frac{\eta^{\mathrm{T}}y - \theta}{\hat{\sigma}_P(y)}.$$

Let $d = \mathrm{tr}(I - P)$ so that

$$W_\theta(y) = \|(I-P)y\|_2^2 + (\eta^{\mathrm{T}}y - \theta)^2 = d\hat{\sigma}_P(y)^2 + (\eta^{\mathrm{T}}y - \theta)^2.$$

Note that τ_θ is independent of W_θ and V.

We next rewrite the quasi-affine inequalities (15) as $U_\theta(y)v + \omega \leqslant \hat{\sigma}_P(y)b$ with

$$v = C\eta, \quad \omega = \omega\{V(y)\} = C\{\theta\eta + V(y)\}.$$

On multiplying both sides by $W_\theta(y)^{1/2}/\hat{\sigma}_P(y)$ we obtain

$$\tau_\theta(y)W_\theta(y)^{1/2}v + \omega\{V(y)\}\{d + \tau_\theta^2(y)\}^{1/2} \leqslant W_\theta(y)^{1/2}b.$$

This is the constraint on the t distribution.

Because $\tau_\theta(y)$ is independent of $\{V(y), W_\theta(y)\}$, its distribution is a truncated $t_{d|\Omega}$, where

$$\Omega(C,b,P,w,v,\theta) = \bigcap_{i=1}^{q} \left\{ t \in \mathbb{R} : t(w^{1/2}v_i) + \omega_i(v)(d+t^2)^{1/2} \leqslant w^{1/2}b_i \right\}. \tag{17}$$

Each individual inequality in this intersection can be solved explicitly, and each yields at most two intervals. □

In practice, the intersection (17) is generally simple, often a single interval.

Remark 1. Following Lemma 4, we can perform selective inference with the quasi-affine constraints of the square-root lasso, where

$$C = -\,\mathrm{diag}(z_S)\,\mathrm{diag}\{(X_S^{\mathrm{T}}X_S)^{-1}\}X_S^{\dagger}, \quad P = P_S, \quad b = -\alpha.$$

To test the hypothesis $H_0 : \beta_{j,S} = 0$, we take $\eta = e_j X_S^{\dagger}$.

3·2. *Inference for σ and debiasing under $\mathbb{M}_{C,b,P}$*

The law $\mathbb{M}_{C,b,P}$ is parametric, and in the context of model selection we have observed y inside the set $Cy \leqslant \hat{\sigma}_P(y)b$. The usual ordinary least squares estimates $X_S^{\dagger}y$ are biased under $\mathbb{M}_{C,b,P} = \mathbb{M}_{C,b,P}(\mu,\sigma^2)$. In this parametric setting, there is a natural procedure for attempting 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 $\mathbb{M}_{C,b,P}$ are $\{\mu/\sigma^2, -1/(2\sigma^2)\}$. Solving the score equations

$$\int_{\mathbb{R}^n} T(z)\, \mathbb{M}_{(C,b,P);(\mu,\sigma^2)}(\mathrm{d}z) - T(y) = 0 \tag{18}$$

for $\{\hat{\mu}(y), \hat{\sigma}^2(y)\}$ corresponds to selective maximum likelihood estimation under $\mathbb{M}_{C,b,P}$. In the orthogonal design and known variance setting, this problem was considered by Reid et al. (2016). In our current setting, solving (18) numerically requires sampling from the constraint set, which is generally nonconvex.

We consider estimation of each parameter separately based on a pseudolikelihood. Unfortunately, in the unknown variance setting, this approach yields estimates for coordinates of μ/σ^2 or σ^2 rather than for μ itself. We propose estimating σ^2 using pseudolikelihood and inserting this value into a distribution analogous to $\mathbb{M}_{C,b,P}$ but with known variance, i.e., the law of $y \sim N(\mu, \sigma^2 I)$, $P\mu = \mu$, with σ^2 known subject to an affine constraint. This approximation is discussed in § 4·2.

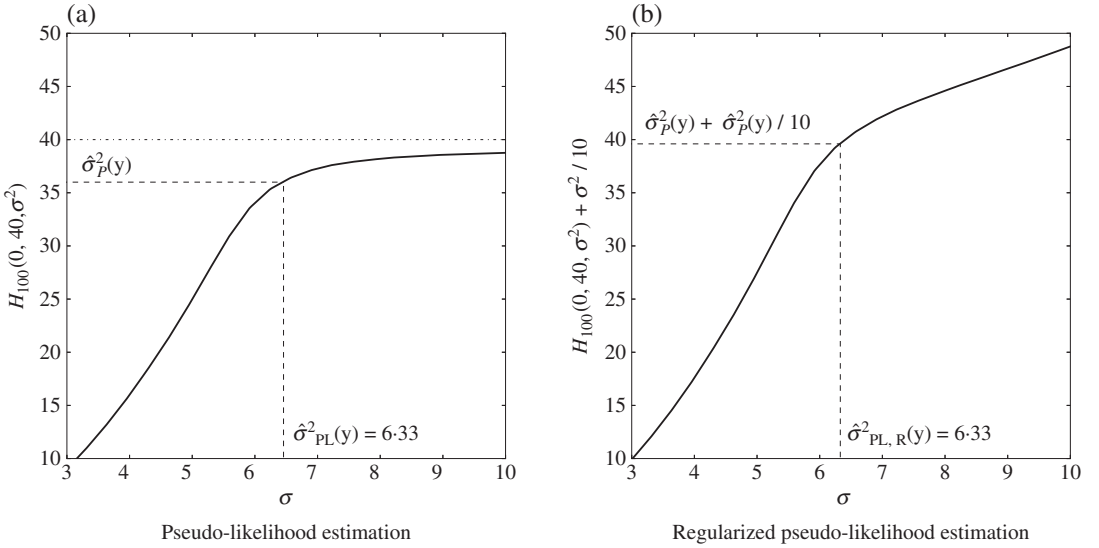

Fig. 1. Estimation of σ^2 based on the pseudolikelihood (19) and the truncation interval with $L = 0$ and $U = 40$; the solid curves are the expectations of the residuals under different values of σ: (a) expectation function of a $\sigma^2\chi^2_{100}$ constrained to $[0, 40]$; (b) the same function plus a regularization term. For the observed value $\hat{\sigma}^2_P(y) = 36$ on 100 degrees of freedom truncated to $[0, 40]$, dashed lines are drawn to find the pseudolikelihood estimator $\hat{\sigma}_{\mathrm{PL}}$ corresponding to $\hat{\sigma}^2_P(y)$; these lines illustrate the root-finding estimator in (20) or (21).

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

$$\|(I - P)y\|_2^2 \mid Py, \quad y \sim \mathbb{M}_{(C,b,P);(\mu,\sigma^2)}. \tag{19}$$

This law depends only on σ^2 and can be used for exact inference about σ^2, though we only use it for estimation. Examining the inequalities, we see that this law is equivalent to $\sigma^2\chi^2_{\mathrm{tr}(I-P)}$ truncated to the interval $[L(Py), U(Py)]$, where

$$L(Py) = \max_{i:\, b_i \geqslant 0} \frac{(Cy)_i}{b_i}, \quad U(Py) = \min_{i:\, b_i \leqslant 0} \frac{(Cy)_i}{b_i}.$$

For $\Omega \subset \mathbb{R}$, let $G_{d,\sigma^2,\Omega}$ denote the law $\sigma^2\chi^2_d$ truncated to Ω, so that $G_{d,\sigma^2,\Omega}(t) = \mathrm{pr}(\chi^2_d \leqslant t/\sigma^2 \mid \chi^2_d \in \Omega/\sigma^2)$.

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

$$\sigma \mapsto H_{\mathrm{tr}(I-P)}\{L(Py), U(Py), \sigma^2\} - \hat{\sigma}^2_P(y), \tag{20}$$

where

$$H_d(L, U, \sigma^2) = \frac{1}{d}\int_{[0,\infty)} t\; G_{d,\sigma^2,[L,U]}(\mathrm{d}t), \quad \hat{\sigma}^2_P(y) = \frac{\|(I - P)y\|^2}{\mathrm{tr}(I - P)};$$

H is readily approximated by sampling from $\sigma^2\chi^2_{\mathrm{tr}(I-P)}$ truncated to $[L(Py), U(Py)]$.

This procedure is illustrated in Fig. 1(a). When observed values of $\hat{\sigma}^2_P(y)$ are near the truncation boundary, the estimate varies quickly with $\hat{\sigma}^2_P(y)$ due to the plateau at the upper limit. To remedy this, we use a regularized estimate of σ under this pseudolikelihood and then apply a bias correction so that when $[L, U] = [0, \infty)$ we recover the usual ordinary least squares estimator.

Specifically, for some θ we obtain a new estimator as the root of

$$\sigma \mapsto H_d(L, U, \sigma^2) + \theta\sigma^2 - (1+\theta)\hat{\sigma}_P^2(y). \tag{21}$$

We call this regularized pseudolikelihood estimate $\hat{\sigma}^2_{\mathrm{RPL}}(y)$. In practice, we have set $\theta = d^{-1/2}$ so that this regularization becomes negligible as the degrees of freedom grows. The regularized estimate can be thought of as the maximum a posteriori estimator from an improper prior on the natural parameter for σ^2. In this case, if $\delta = 1/(2\sigma^2)$ is the natural parameter, the prior has density proportional to $\delta\,\mathrm{d}\theta$. As $H_d(0, \infty, \sigma^2) = \sigma^2$, it is clear that in the untruncated case we recover the usual ordinary least squares estimator $\hat{\sigma}_P(y)$.

4. Applications

4·1. *Comparison of estimators*

We now study the accuracy of our estimator $\hat{\sigma}_{\mathrm{RPL}}$ of σ by comparing it with other estimators in the lasso literature: the ordinary least squares estimator in the selected model, where $\hat{\sigma}_S = \|(I - P_S)y\|/(n - |S|)^{1/2}$ with S being the active set; the scaled lasso estimator of Sun & Zhang (2012); and the minimum crossvalidation estimator of Reid et al. (2016), based on the residual sum of squares of the lasso solution with λ selected by crossvalidation.

We consider a high-dimensional setting, generating design matrices X of dimension 1000×2000 with each entry normally distributed and with columns having pairwise correlation equal to 0·3. The columns were also normalized to have length 1. We then generated

$$y = X\beta + \epsilon, \quad \beta = (\underbrace{\pm\Delta, \ldots, \pm\Delta}_{s}, 0, \ldots, 0), \quad \epsilon \sim N(0, \sigma^2 I). \tag{22}$$

The sparsity was set to $s = 40$, and each coefficient had equal magnitude, $\Delta = 8$, but a random sign. The noise level $\sigma = 3$ is considered unknown. The parameter κ in (4) was set to 0·8. With these settings, the square-root lasso discovered a superset of the 40 nonzero coefficients, with a success rate of approximately 30%. Since most of the time we do not screen, the model is sometimes misspecified. But even in these cases, the variance estimator $\hat{\sigma}_{\mathrm{RPL}}$ is consistent with the model-specific variance

$$\sigma_S = \left(\sigma^2 + \left\|X_S\hat{\beta}_S - X\beta\right\|^2\right)^{1/2}.$$

We compare the performance of the estimators by considering the ratio $\hat{\sigma}(y)/\sigma_S$, where σ_S is the usual variance estimate obtained using the selected variables S, evaluated on an independent copy of data drawn from the same distribution. This is the variance one would expect to see in long-run sampling if fitting an ordinary least squares model with variables S under the true data-generating distribution.

Figure 2(a) shows that our estimator outperforms the ordinary least squares estimator, the scaled lasso estimator and the minimum crossvalidation estimator. Although we did not establish consistency of our estimator, it has only a small upward bias, which makes it slightly conservative. Furthermore, the robustness of $\hat{\sigma}_{\mathrm{PL}}$ is demonstrated in Fig. 2(b), where the active set S fails to include all nonzero variables. In the case of partial recovery of the true coefficient set, we see that $\hat{\sigma}_{\mathrm{PL}}$ is quite close to σ_S; here, part of the noise in our selected model comes from missing some of the true signals. Our estimator beats all the other estimators in this case. In particular, the ordinary least squares estimator consistently underestimates the variance, which could lead to inflated test scores and more false discoveries. On the other hand, minimum crossvalidation seeks to estimate σ^2 instead of σ_S^2, which results in a downward bias.

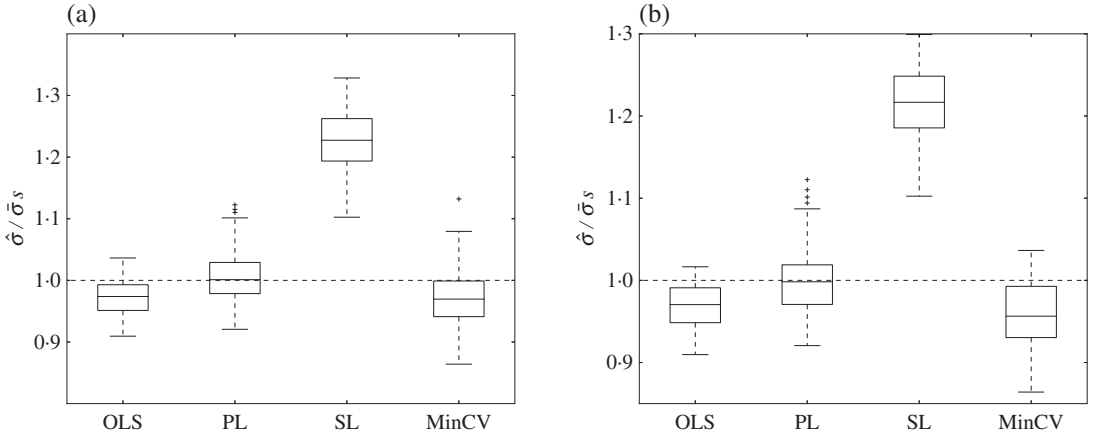

Fig. 2. Comparisons of different estimators for σ_S: OLS, the usual ordinary least squares estimator for selected variables S; PL, the pseudolikelihood estimate $\hat{\sigma}_{\mathrm{PL}}(y)$; SL, the scaled lasso estimate of scale; MinCV, the minimum crossvalidation estimate of Reid et al. (2016). (a) The ratio $\hat{\sigma}/\sigma_S$ in all 100 simulations, making no distinction according to whether S contains all nonzero coefficients; (b) the ratio $\hat{\sigma}/\sigma_S$ in only those simulations where the selected set S does not include all the nonzero coefficients. In each panel the dashed line represents the ratio of 1 for an unbiased estimator.

In fact, when (y_i, X_i) are independent draws from a fixed Gaussian distribution, then for any S the model M_S is correctly specified in the sense that the law of $y \mid X_S$ belongs to the family M_S. In this setting the quantity σ_S^2 is an asymptotically correct estimator $\sigma_S^2 = \mathrm{var}(y_i \mid x_{i,S})$. Hence, we see that the pseudolikelihood estimator may be considered a reasonable estimator when the square-root lasso does not actually screen.

4·2. *An approximation to* $\mathbb{M}_{C,b,P}$ *and non-Gaussian noise*

The selective distribution $\mathbb{Q}_{S,z_S}$ derived from some $\mathrm{pr}_S \in M_S$ is used for inference about the parameters β_S. Let η be the normalized linear functional for testing $H_0 : \beta_{j,S} = \theta$; then for any value of θ this distribution is restricted to the intersection of the sphere of radius $\|y - \theta X_j\|_2$ with the affine space $\{z : X_{S\setminus j}^{\mathrm{T}} z = X_{S\setminus j}^{\mathrm{T}} y\}$. Call this set $\Theta = \Theta(\|P_{S\setminus j}(y - \theta\eta)\|_2, X_{S\setminus j}^{\mathrm{T}} y)$. The restriction of $\mathbb{Q}_{S,z_S}$ to Θ is the law pr_S restricted to the intersection of Θ with the selection event. For $|S|$ that is not large relative to n, by Poincaré's limit (Diaconis & Freedman, 1987), the law of $\eta^{\mathrm{T}} y - \theta$ under pr_S restricted to Θ is close to a Gaussian distribution with variance $\|P_{S\setminus j}(y - \theta\eta)\|_2^2/(n - |S| + 1)$. Hence, we could approximate its distribution under $\mathbb{Q}_{S,z_S}$ by a truncated Gaussian distribution. Furthermore, we can approximate the variance by $\hat{\sigma}_{\mathrm{RPL}}(y)^2$. We summarize the Gaussian approximation in the following remark.

Remark 2. Suppose we are interested in testing the hypothesis $H_0 : \eta^{\mathrm{T}}\mu = \theta$ in the family $\mathbb{M}_{C,b,P}$, for some $\eta \in \mathrm{row}(C)$. We propose using the distribution

$$\mathcal{L}\big\{\eta^{\mathrm{T}} z - \theta \;\big|\; (P - \eta\eta^{\mathrm{T}})z, Cz \leqslant \hat{\sigma}_{\mathrm{RPL}}(y) b\big\}, \quad z \sim N\{\mu, \hat{\sigma}_{\mathrm{RPL}}^2(y) I\},$$

with $\hat{\sigma}_{\mathrm{RPL}}(y)$ interpreted as a constant plug-in estimate. We condition on $(P - \eta\eta^{\mathrm{T}})z$, as we have assumed $P\mu = \mu$ in defining $\mathbb{M}_{C,b,P}$ and this is a sufficient statistic for the unknown parameter $(P - \eta\eta^{\mathrm{T}})\mu$.

We validated the above approximation through simulations. We generated design matrices of dimension 150×200. We again normalized the columns and set the pairwise column correlations

Table 1. *Mean and standard deviation of the true coverage rate averaged over* 100 *replicates; levels are the nominal levels of the intervals*

(a) Coverage using approximations to $M_{C,b,P}$			(b) Coverage under heavy-tailed noises		
Level	$\overline{\text{TCR}}$	SD($\overline{\text{TCR}}$)	Level	$\overline{\text{TCR}}$	SD($\overline{\text{TCR}}$)
85%	86·2%	1·5%	85%	87·6%	1·3%
90%	89·9%	1·5%	90%	92·6%	1·1%
95%	94·5%	1·1%	95%	94·1%	1·3%
97%	97·5%	0·6%	97%	95·6%	1·3%

$\overline{\text{TCR}}$, mean of the true coverage rate; SD($\overline{\text{TCR}}$), standard deviation of the true coverage rate.

to be 0·3. We then generated y from (22) with sparsity $s = 10$ and with the value of each nonzero coefficient being $\Delta = 6$. The signs of the coefficients were fixed to be positive; $\sigma = 1$ is assumed unknown. We form level-$(1-\alpha)$ confidence intervals for the coefficients selected by the square-root lasso. Each instance of the data-generating mechanism produces a true coverage rate, or 1 minus the false coverage rate of Benjamini & Yekutieli (2005). We report the mean and standard deviation of the true coverage rate over 100 replications in Table 1(a).

The theory and methods developed in this work involve normality assumptions on the data. However, the limits of such assumptions can be probed in simulations. In particular, we generate y from (22), except that now the ϵs are generated from a t distribution with five degrees of freedom. The design matrices and other parameters such as s, Δ and σ are the same as above. We compare the coverage of our methods with heavy-tailed noises in Table 1(b). We see that although the coverages are not as exact as in Table 1(a), they are close to the nominal levels. This indicates that we can apply the proposed methods to data even when the normality assumptions are violated.

4·3. *Regression diagnostics*

Recall that the scaled residual vector $U_{-S}(y)$ in (13) is ancillary under the laws $m_S \in M_S$ and $\tilde{\mathbb{Q}}_{S,z_S}$; $U_{-S}(y)$ follows a uniform distribution on the intersection of the n-dimensional unit sphere with the subspace determined by $I - P_S$, truncated by the observed constraints (12). As $U_{-S}(y)$ 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 $\|\hat{U}_{-S}(y)\|_\infty$.

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

$$F_{G|G\cup S}(y) = \frac{\|(P_{G\cup S} - P_S)y\|_2^2/|G|}{\|(I - P_{G\cup S})y\|_2^2/(n - |G \cup S|)} = \frac{\|(P_{G\cup S} - P_S)U_{-S}(y)\|_2^2/|G|}{\|(I - P_{G\cup S})U_{-S}(y)\|_2^2/(n - |G \cup S|)}.$$

Therefore, a selectively valid test of $H_0 : \beta_{G|G\cup S} = 0$ can be constructed by sampling U_{-S} under its null distribution and comparing the observed F-statistic to this reference distribution. Details of these diagnostics are a potential area of further research.

4·4. *Applications to false discovery rate control*

Based on the truncated t distribution derived in Theorem 1, it is easy to construct tests that control the selective Type I error. In this subsection, we attempt to use our p-values for

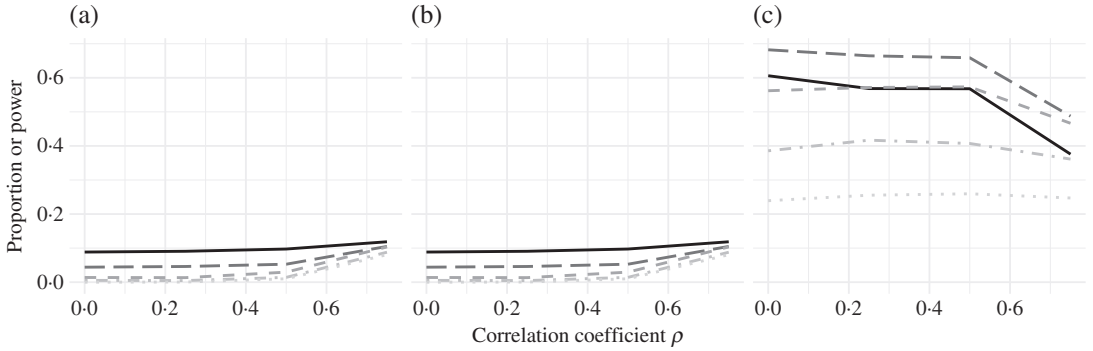

Fig. 3. False discovery rate control and power using the Benjamini-Hochberg procedure on our adjusted p-values for variables selected by the square-root lasso: (a) directional false discovery proportion; (b) full model false discovery proportion; (c) full model power. To ensure fairness of the comparison with Barber & Candès (2018), we use the data-generating mechanism in § 5 of their paper, where $n = 2000$, $p = 2500$, $k = 30$, and the rows of X are generated from $N(0, \Sigma)$. The covariance $\Sigma_{ij} = \rho^{|i-j|}$ has a tapered structure, and we vary ρ along the horizontal axis of each panel. We also vary κ in (4); solid, long-dashed, dashed, dot-dashed and dotted lines of decreasing darkness correspond to $\kappa = 0{\cdot}6, 0{\cdot}7, 0{\cdot}8, 0{\cdot}9$ and $0{\cdot}1$, respectively. Plots show the average of 200 independent replicates.

multiple hypothesis testing purposes. We apply the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) to our p-values and compare this approach with the method of Barber & Candès (2018). The results are shown in Fig. 3. Our procedure is relatively robust with respect to the choice of κ in both full model false discovery rate control and full model power across different correlations ρ. For all choices of κ, our procedure controls the false discovery rate at $0{\cdot}2$ across all the correlation coefficients ρ. Our method enjoys an approximately 20% increase in power over that of Barber & Candès (2018, Fig. 1). We cannot hope that this advantage will persist across all choices of κ. If κ is too large, then the selection algorithm will be conservative, leaving out true features. If κ is too small, then we may select too many features, and the selection event will leave little leftover information (Fithian et al., 2017). Methods of fine tuning the parameter κ are beyond the scope of this paper, but we recommend using κ within the range of $0{\cdot}7$ to $0{\cdot}8$ in practice. In the scenario of Barber & Candès (2018), this range of κ achieved superior performance across all feature correlations ρ.

Acknowledgement

Tian and Taylor were supported by the U.S. National Science Foundation. Loftus was supported by the Alan Turing Institute during part of this work.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of all the lemmas.

References

Barber, R. F. & Candès, E. J. (2018). A knockoff filter for high-dimensional selective inference. *arXiv:* 1602.03574v3.

Belloni, A., Chernozhukov, V. & Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98**, 791–806.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Benjamini, Y. & Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Am. Statist. Assoc.* **100**, 71–81.

DIACONIS, P. & FREEDMAN, D. A. (1987). A dozen de Finetti-style results in search of a theory. *Ann. Inst. Henri Poincaré Prob. Statist.* **23**, 397–423.

FITHIAN, W., SUN, D. & TAYLOR, J. (2017). Optimal inference after model selection. *arXiv:* 1410.2597v4.

LEE, J. D., SUN, D. L., SUN, Y. & TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44**, 907–27.

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. & YU, B. (2012). A unified framework for high-dimensional analysis of MM-estimators with decomposable regularizers. *Statist. Sci.* **27**, 538–57.

REID, S., TIBSHIRANI, R. J. & FRIEDMAN, J. H. (2016). A study of error variance estimation in lasso regression. *Statist. Sinica* **26**, 35–67.

SUN, T. & ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–98.

TAYLOR, J. E., LOFTUS, J. R. & TIBSHIRANI, R. J. (2016). Inference in adaptive regression via the Kac–Rice formula. *Ann. Statist.* **44**, 743–70.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Statist.* **7**, 1456–90.

TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. & TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Am. Statist. Assoc.* **111**, 600–20.

[*Received on* 30 *November* 2016*. Editorial decision on* 18 *March* 2018]