Post-selection inference for ℓ_1-penalized likelihood models

Jonathan TAYLOR and Robert TIBSHIRANI*

Departments of Statistics and Biomedical Data Science, Stanford University, Stanford CA 94305, U.S.A.

Abstract: We present a new method for post-selection inference for ℓ_1 (lasso)-penalized likelihood models, including generalized regression models. Our approach generalizes the post-selection framework presented in Lee et al. (2013). The method provides *P*-values and confidence intervals that are asymptotically valid, conditional on the inherent selection done by the lasso. We present applications of this work to (regularized) logistic regression, Cox's proportional hazards model, and the graphical lasso. We do not provide rigorous proofs here of the claimed results, but rather conceptual and theoretical sketches. *The Canadian Journal of Statistics* 46: 41–61; 2018 © 2017 Statistical Society of Canada

Résumé: Les auteurs présentent une nouvelle méthode d'inférence post-sélection pour les modèles de vraisemblance avec une pénalité ℓ_1 (lasso). Leur approche généralise le cadre d'inférence post-sélection de Lee et coll. (2013). Leur méthode génère des p-values et des intervalles de confiance qui sont asymptotiquement valides conditionnellement à la sélection inhérente au lasso. Les auteurs présentent une application de ces résultats à la régression logistique (régularisée), au modèle à risques proportionnels de Cox et au lasso graphique. Ils ne présentent pas de preuves rigoureuses des résultats avancés, mais plutôt une esquisse conceptuelle et théorique. *La revue canadienne de statistique* 46: 41–61; 2018 © 2017 Société statistique du Canada

1. INTRODUCTION

Significant recent progress has been made in the problem of inference after selection for Gaussian regression models. In particular Lee et al. (2013) derives closed form *P*-values and confidence intervals, after fitting the lasso with a fixed value of the regularization parameter, and Taylor et al. (2014) provides analogous results for forward stepwise regression and least angle regression (LAR). In this article we derive a simple and natural way to extend these results to ℓ_1-penalized likelihood models, including generalized regression models such as (regularized) logistic regression and Cox's proportional hazards model.

Formally the results described here are contained in Tian & Taylor (2015) in which the authors consider the same problems but having added noise to the data before fitting the model and carrying out the selection. Besides expanding on the GLM case, considered only briefly in Tian & Taylor (2015), the novelty in this work is the second-stage estimator, which is asymptotically equivalent to the post-LASSO MLE, and overcomes some problems encountered with a second-order remainder. And unlike the proposals in Tian & Taylor (2015) the estimator proposed here does not require MCMC sampling and is computable in closed form. Our estimator comprises a single step of Newton–Raphson (or equivalently, Fisher scoring) in the "selected model" after having fit the LASSO. This is discussed further in Section 3. We note that one-step estimators are commonly used in semi-parametric inference: see, for example, Bickel et al. (1993).

---

* *Author to whom correspondence may be addressed.*
 E-mail: tibs@stanford.edu

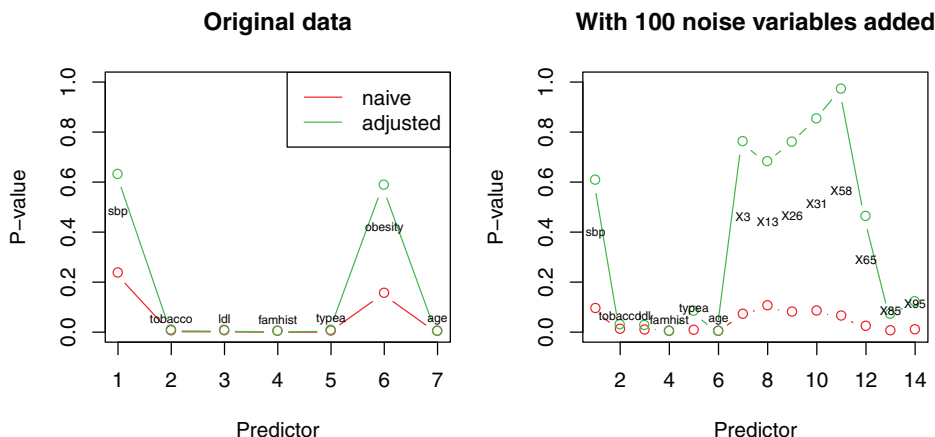**Original data**   **With 100 noise variables added**

FIGURE 1: South African heart disease data. *P*-values from naive and selection-adjusted approaches, for original data (left) and data with 100 additional noise predictors (right). Each model was chosen by lasso-penalized logistic regression, choosing the tuning parameter by cross-validation. The small *P*-values in the left panel are all below 0.001.

In this article we do not provide rigorous proofs of the claimed results, but rather theoretical and conceptual sketches, together with numerical evidence. We are confident that rigorous proofs can be given (with appropriate assumptions) and plan to report these elsewhere. We also note the strong similarity between our one-step estimator and the "debiased lasso" construction of Bühlmann (2013), van de Geer et al. (2013), and Javanmard & Montanari (2014), Zhang & Zhang (2014). This connection is detailed in Remark A of this article.

Figure 1 shows an example, the South African heart disease data. This is a retrospective sample of 463 males in a heart-disease high-risk region of the Western Cape, South Africa. The outcome is binary—coronary heart disease—and there are nine predictors. We applied lasso-penalized logistic regression, choosing the tuning parameter by cross-validation. The left panel shows the standard (naive) *P*-values and the post-selection *P*-values from our theory, for the predictors in the active set. As the sample size is large compared to the number of predictors, the unadjusted and adjusted *P*-values are only substantially different for two of the predictors. On the right we have added 100 independent standard Gaussian predictors (labelled $X1, X2 \dots X100$) to examine the effects of selection. Now the naive *P*-values are unrealistically small for the noise variables, whereas the adjusted *P*-values are appropriately large. Although our focus is on selection via ℓ_1-penalization, a similar approach can likely be applied to forward stepwise methods for likelihood models, and in principle, least angle regression (LAR).

An outline of this article is as follows. Section 2 reviews post-selection inference for the lasso in the Gaussian regression model and introduces our proposal for more general (non-Gaussian) generalized linear models. In Section 3 we give an equivalent form of the proposal, one that applies to general likelihood models, for example the graphical lasso. We give a rough argument for the asymptotic validity of the procedure. Section 4 reports a simulation study of the methods. In Section 5 we show an example of the proposal applied to Cox's model for survival data. The graphical lasso is studied in Section 6. We end with a discussion in Section 7.

2. POST-SELECTION INFERENCE FOR GENERALIZED REGRESSION MODELS

Suppose that we have data (x_i, y_i), $i = 1, 2, \dots N$ consisting of features $x_i = (x_{i1}, x_{i2}, \dots x_{ip})$ and outcomes y_i, $i = 1, 2, \dots N$. Let $X = \{x_{ij}\}$ be the $N \times p$ data matrix. We consider a generalized regression model with linear predictor $\eta = \alpha + x^\top \beta$ and log-likelihood $\ell(\alpha, \beta)$. Our objective

function has the form

$$J(\alpha, \beta) = -\ell(\alpha, \beta) + \lambda \cdot \sum_1^p |\beta_j|. \tag{1}$$

Let $\hat{\alpha}$, $\hat{\beta}$ be the minimizers of $J(\alpha, \beta)$. We wish to carry out post-selection inference for some functional $\gamma^\top \beta$. For example γ might be chosen so that $\gamma^\top \beta$ is the population partial regression coefficient for the jth predictor.

As a leading example we consider the logistic regression model specified by

$$\pi = E(Y|x); \ \log \pi/(1 - \pi) = \alpha + x^\top \beta; \tag{2}$$

$$\ell(\alpha, \beta) = \sum [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)].$$

Having fit this model using a fixed value of λ we carry out post-selection inference, as in the heart disease example above.

The reader may well ask: "partial regression coefficient with respect to what?" that is, what other covariates are we going to control for? In this article we follow the selected model framework described in Fithian, Sun, & Taylor (2014) so that having observed M, the sparsity pattern of $\hat{\beta}$ as returned by the LASSO, we carry out selective inference for linear functionals of $\beta_M \in \mathbb{R}^M$ under the assumption that the model (2) is correct with parameter β^* satisfying $\beta^*_{-M} = 0$. That is we carry out selective inference under the assumption that the LASSO has screened successfully, at least approximately.

There are various ways we might modify this model, though we only consider mainly the parametric case here. For instance we might assume that y is conditionally independent of X_{-M} given X_M, but not assume the correctness of the logistic link function. In this case the covariance matrix of our limiting Gaussian distribution (described below) is not correct, and the asymptotic theory in Section 3 should be modified by using a consistent estimator of the covariance matrix. Alternatively we might wish to be robust to the possibility that X_{-M} may have some effect on our sampling distribution, which would also change the limiting Gaussian distribution that we use for inference. For this reason the results in Section 4.2 of Tian & Taylor (2015) use bootstrap or jackknife to estimate the appropriate covariance in the limiting Gaussian distribution. A short discussion of how this can be done is given in Section 3.4. As robustness to various mis-specifications are important issues, in this article we focus mainly on the simpler case of providing inference for parameters of (2) under the assumption that the model chosen by the LASSO screens, that is, has found a superset of the true variables.

2.1. Review of the Gaussian Case

For background we first review the Gaussian case $y \sim N(\mu, I \cdot \sigma^2)$, developed in Lee et al. (2013). We denote the selected model by M with sign vector s. Assuming the columns of X are in general position, the KKT conditions Tibshirani (2013) state that $\{\hat{M}, \hat{s}\} = \{M, s_M\}$ if and only if there exists $\hat{\beta}_M \in \mathbb{R}^M$ and $u_{-M} \in \mathbb{R}^{-M}$ satisfying

$$\begin{aligned} X_M^\top (X_M \hat{\beta}_M - y) + \lambda s_M &= 0 \\ X_{-M}^\top (X_M \beta_M - y) + \lambda u_{-M} &= 0 \\ \mathrm{sign}(\hat{\beta}_M) &= s_M \\ \|u_{-M}\|_\infty &\leq 1. \end{aligned} \tag{3}$$

This allows us to write the set of responses y that yield the same M and s in the polyhedral form

$$\{Ay \le b\}, \tag{4}$$

where the matrix A and vector b depend on X and the selected model, but not on y. Let P_M be the orthogonal projector onto the model subspace. Due to the special form of the LASSO optimization problem, it turns out that the rows of A (and b) can be partitioned so that we can rewrite the above as

$$\{A_1 \hat{\beta}_M(y) \le b_1, A_2(I - P_M)y \le b_2\}, \tag{5}$$

where $\hat{\beta}_M(y) = (X_M^\top X_M)^{-1} X_M^\top y$ are the usual OLS estimators of β_M and $(I - P_M)y$ are the usual OLS residuals.

This result can be used to make conditional inferences about any linear functional $\gamma^\top \mu$, which we assume satisfies $P_M \gamma = \gamma$. This assumption is roughly equivalent to assuming that we are interested in a linear functional of $P_M \mu$. By conditioning on $P_\gamma^\perp(y)$ we obtain the exact result based on truncated Gaussian distribution with truncation limits $\mathcal{V}^-, \mathcal{V}^+$:

$$F_{\gamma^\top \mu, \sigma^2 ||\gamma||^2}^{\mathcal{V}^-, \mathcal{V}^+} \left(\gamma^\top y \right) \mid \{Ay \le b\} \sim \mathrm{U}(0, 1). \tag{6}$$

Expressions for A, b and the limits $\mathcal{V}^-, \mathcal{V}^+$ are given in Lee et al. (2013) and are reproduced here in the Appendix. The relation (5) implies that the result (6) holds even if we condition only on $(P_M - P_\gamma)y$, that is, the variation of y within the model. This follows as the second condition in (5) is independent of the first condition, and is fixed after conditioning on $P_\gamma^\perp(y)$. The difference between these two conditional distributions really depends on which model we are interested in. We refer to conditioning on $P_\gamma^\perp y$ as inference in the saturated model, that is, the collection of distributions

$$\left\{ N(\mu, \sigma^2 I) : \mu \in \mathbb{R}^n, \sigma^2 > 0 \right\}.$$

We refer to conditioning on $(P_M - P_\gamma)y$ as inference in the selected model. Formally speaking we define the selected model as follows: given a subset of variables E, the selected model corresponding to E is the collection of distributions

$$\left\{ N(X_E \beta_E, \sigma^2 I) : \beta_E \in \mathbb{R}^E, \sigma^2 > 0 \right\}. \tag{7}$$

This distinction is elaborated on in Lee et al. (2013). In this article we only consider inference under (7) where the subset of variables E are those chosen by the LASSO. In principle, however, a researcher can add or delete variables from this set at will if they make their decisions based only on the set of variables chosen by the LASSO. This changes the distributions for inference, meaning that the analog of (6) may no longer be the correct tool for inference.

2.2. Extension to Generalized Regression Models

In this section we make a parallel between the Gaussian case and the generalized linear model setting. This parallel should be useful to statisticians familiar with the usual iteratively reweighted least squares (IRLS) algorithm to fit the unpenalized logistic regression model. As this parallel may be useful the formal justification is given in Section 3. The method used to solve the optimization problem is unrelated to the results presented in Section 3.

A common strategy for minimizing (1) is to express the usual Newton–Raphson update as an IRLS step, and then replace the weighted least squares step by a penalized weighted least squares

procedure. For simplicity we assume $\alpha = 0$ below, though our formal justification in Section 3 will allow for an intercept as well as other unpenalized features.

In detail, recalling that ℓ is the log-likelihood, we define

$$W = W(\beta) = -\left(\frac{\partial^2 \ell}{\partial \eta \partial \eta^\top}\right)\bigg|_{\eta = X\beta}$$

and

$$z = z(\beta) = X\beta + W^{-1}\left(\frac{\partial \ell}{\partial \eta}\right)\bigg|_{\eta = X\beta}.$$

Of course, in the Gaussian case, $W = I$ and $z = y$.

In this notation the Newton–Raphson step (in the unpenalized regression model) from a current value β_c can be expressed as

$$\text{minimize}_\beta \frac{1}{2}(z(\beta_c) - X\beta)^\top W(\beta_c)(z(\beta_c) - X\beta). \tag{8}$$

In the ℓ_1-penalized version each minimization has an ℓ_1 penalty attached.

To minimize (1), IRLS proceeds as follows:

 (i) Initialize $\hat{\beta} = 0$
 (ii) Compute $W(\hat{\beta})$ and $z(\hat{\beta})$ based on the current value of $\hat{\beta}$
(iii) Solve

$$\text{minimize}_\beta \frac{1}{2}(z - X\beta)^\top W(z - X\beta) + \lambda \cdot \sum |\beta_j|.$$

(iv) Repeat steps 2 and 3 until $\hat{\beta}$ does not change more than some pre-specified threshold.

In logistic regression, for example, the specific forms of the relevant quantities are

$$z = \hat{\alpha} + X\hat{\beta} + (y - \hat{\pi}) \cdot \text{diag}[1/(\hat{\pi}(1 - \hat{\pi}))]$$
$$W = \text{diag}(\hat{\pi}_i(1 - \hat{\pi}_i)). \tag{9}$$

Another example is Cox's proportional hazards for censored survival data. The partial likelihood estimates can again be found via an IRLS procedure. More generally both of these examples are special cases of the one-step estimator described in Section 3 below. The details of the adjusted dependent variable z and weights w can be found, for example, in Hastie & Tibshirani (1990) (Chapter 8, pp. 213–214). We give examples of both of these applications later.

How do we carry out post-selection inference in this setting? Our proposal is to treat the final iterate as a weighted least squares regression, and hence use the approximation

$$z \sim N(\mu, W^{-1}). \tag{10}$$

Using this idea we simply apply the polyhedral lemma to the region $\{Az \leq b\}$ (see the Appendix). A potential problem with this proposal is that A and b depend on $\hat{\beta}$ and hence on y. As a result the region $Az \leq b$ does not exactly correspond to the values of the response vector y yielding the same active set and signs as our original fit. The other obvious problem is that of course z is not actually normally distributed. Despite these points we provide evidence in Section 3 that this procedure yields asymptotically correct inferences.

2.3. Details of the Procedure

Suppose that we have iterated the above procedure until we are at a fixed point. The "active" block of the stationarity conditions has the form

$$X_M^\top W(z - X_M \hat{\beta}_M) = \lambda s_M, \tag{11}$$

where $W = W(\hat{\beta}_M)$, $z = z(\hat{\beta}_M)$. Solving for $\hat{\beta}_M$ yields

$$\hat{\beta}_M = (X_M^\top W X_M)^{-1}(X_M^\top W z - \lambda s_M).$$

Thinking of z as analogous to y in the Gaussian case, this equality can be re-expressed as

$$\bar{\beta}_M \equiv (X_M^\top W X_M)^{-1} X_M^\top W z = \hat{\beta}_M + \lambda (X_M^\top W X_M)^{-1} s_M. \tag{12}$$

Note that $\bar{\beta}_M$ solves

$$X_M^\top W(z - X_M \bar{\beta}_M) = 0.$$

This last equation is "almost" the stationarity conditions of the unpenalized MLE for the logistic regression using only the features in M. The only difference is that above, the W and z are evaluated at $\hat{\beta}_M$ instead of $\bar{\beta}_M$.

Ignoring this discrepancy for the moment, recall that the active block of the KKT conditions in the Gaussian case can be expressed in terms of the usual OLS estimators (5). This suggests the correct analog of the "active" constraints:

$$\left\{ y : \text{sign}(\bar{\beta}_M(y) - \lambda(X_M^\top W X_M)^{-1} s_M) = s_M \right\}. \tag{13}$$

Let us take a closer look at $\bar{\beta}_M$:

$$\begin{aligned}
\bar{\beta}_M &= (X_M^\top W X_M)^{-1}(X_M^\top W z) \\
&= (X_M^\top W X_M)^{-1}\left(X_M^\top W X_M \hat{\beta}_M + \left.\frac{\partial}{\partial \beta_M}\ell_M(\beta_M)\right|_{\beta_M = \hat{\beta}_M} \right) \\
&= \hat{\beta}_M + (X_M^\top W X_M)^{-1}\left.\frac{\partial}{\partial \beta_M}\ell_M(\beta_M)\right|_{\beta_M = \hat{\beta}_M} \\
&= \hat{\beta}_M + \lambda(X_M^\top W X_M)^{-1} s_M,
\end{aligned}$$

where

$$\ell_M(\beta_M) \equiv \ell\begin{pmatrix} \beta_M \\ 0 \end{pmatrix}$$

is the log-likelihood of the selected model and

$$I_M(\hat{\beta}_M) = X_M^\top W X_M = X_M^\top W(\hat{\beta}_M) X_M$$

is its Fisher information matrix evaluated at $\hat{\beta}_M$. We see that $\bar{\beta}_M$ is defined by one Newton–Raphson step in the selected model from $\hat{\beta}_M$.

If we had not used the data to select variables M and signs s_M, then assuming the model with variables M is correctly specified, as well as standard assumptions on X (Bunea, 2008), $\bar{\beta}_M$ would be asymptotically Gaussian centred around β_M^* with approximate covariance $(X_M^\top W X_M)^{-1}$. This approximation is of course the usual one used in forming Wald tests and confidence intervals in generalized linear models.

3. A MORE GENERAL FORM AND AN ASYMPTOTIC JUSTIFICATION

We assume that p is fixed, that is, our results do not apply in the high-dimensional regime where $p \to \infty$. To state our main result we begin by considering a general lasso-penalized problem. Given a log-likelihood $\ell(\beta)$, denote an ℓ_1-penalized estimator by

$$\hat{\beta} = \hat{\beta}_\lambda = \mathrm{argmin}_\beta [-\ell(\beta)] + \lambda \|\beta\|_1. \tag{14}$$

On the event $\{(\hat{M}, s_{\hat{M}}) = (M, s_M)\}$, the active block of the KKT conditions is

$$\left. \frac{\partial}{\partial \beta_M} \ell_M(\beta_M) \right|_{\beta_M = \hat{\beta}_M} = \lambda s_M,$$

where ℓ_M is the log-likelihood of the submodel M. The corresponding one-step estimator is

$$\begin{aligned}
\bar{\beta}_M &= \hat{\beta}_M + \lambda I_M(\hat{\beta}_M)^{-1} s_M \\
&= \hat{\beta}_M + I_M(\hat{\beta}_M)^{-1} \left. \frac{\partial}{\partial \beta_M} \ell_M(\beta_M) \right|_{\beta_M = \hat{\beta}_M},
\end{aligned} \tag{15}$$

where $I_M(\hat{\beta}_M)$ is the $|M| \times |M|$ observed Fisher information matrix of the submodel M evaluated at $\hat{\beta}_M$.

In the previous section we noted that $\bar{\beta}_M$ "almost" solves the KKT conditions for the unpenalized logistic regression model. We further recognized it as a one step estimator with initial estimator $\hat{\beta}_M$ in the logistic regression model. In the context (14) we have directly defined $\bar{\beta}_M$ as a one-step estimator with the initial estimator $\hat{\beta}_M$. As long as λ is selected so that $\hat{\beta}_M$ is $\sqrt{n}$ consistent (usually satisfied by taking $\lambda \propto n^{1/2}$ at least in the fixed p setting considered here) the estimator $\bar{\beta}_M$ would typically have the same limiting distribution as the unpenalized MLE in the selected model if we had not used the data to choose the variables to be included in the model. That is if we had not selected the variables based on the data, standard asymptotic arguments yield

$$\bar{\beta}_M \approx N\left(\beta_M^*, I_M(\hat{\beta}_M)^{-1} \right), \tag{16}$$

where $I_M(\hat{\beta}_M)^{-1}$ is the "plug-in" estimate of the asymptotic covariance of $\bar{\beta}_M$, with the population value being $E_F[I_M(\beta_M^*)]^{-1}$. Implicit in this notation is that $I_M = I_{M,n}$, that is, the information is based on a sample of size n from some model. We specify this model precisely in Section 3.1 below.

However selection with the LASSO has imposed the "active" constraints, that is, we have observed the following event is true

$$\left\{ \mathrm{diag}(s_M) \left[\bar{\beta}_M - I_M(\hat{\beta}_M)^{-1} \lambda s_M \right] \geq 0 \right\}, \tag{17}$$

as well as some "inactive" constraints that we return to shortly. Selective inference Fithian, Sun, & Taylor (2014), Taylor & Tibshirani (2015) modifies the pre-selection distribution by conditioning

on the LASSO having chosen these variables and signs, that is, by conditioning on this information we have learned about the data.

3.1. Specification of the Model

As the objective function involves a log-likelihood there is some parametric family of distributions that is natural to use for inferential purposes. In the generalized linear model setting these distributions are models for the laws $y_i|X$. Combining this with a marginal distribution for X yields a full specification of the joint law of (y, X). In our only result below we consider the X_i's to be IID draws from some distribution F and $y_i|X$ to be independently drawn according to the log-likelihood corresponding to the generalized linear model setting. In this case our model is specified by a pair (β, F) and we can now consider asymptotic behaviour of our procedure sending $n \to \infty$. Similarly for p fixed and any $M \subset \{1, \ldots, p\}$ our selected model is specified by the pair (β_M, F) and we can consider similar asymptotic questions. As is often the case the most interesting asymptotic questions are local alternatives in which β_M itself depends on n, typically taking the form $\beta_{M,n}^* = n^{-1/2}\theta_M^*$. These assumptions are similar to those studied in Bunea (2008). In local coordinates (16) could more properly be restated as

$$\bar{\theta}_M = n^{1/2}\bar{\beta}_M \approx N\left(\theta_M^*, nI_M(\hat{\beta}_M)^{-1}\right), \tag{18}$$

where $nI_M(\hat{\beta}_M)^{-1}$ will have non-zero limit

$$\lim_{n\to\infty} E_F[I_M(n^{1/2}\theta_M^*)/n]^{-1}.$$

For example, in the Bernoulli case (binary Y), for any $M \subset \{1, \ldots, p\}$ and sample size n our selected model is therefore parametrized by (θ_M^*, F) where features are drawn IID according to F and, conditionally on X_i we have

$$y_i|X_i \sim \text{Bernoulli}(\pi(n^{-1/2}\theta_M^*)).$$

3.2. Asymptotics of the One-Step Estimator

In this section we lay out a description of the limiting conditional distribution of the one-step estimator in the logistic case. Under our local alternatives, in the selected model the data generating mechanism is completely determined by the tuple (n, θ_M^*, F) where F is the distribution of the features X and $y|X$ is assumed to follow the parametric logistic regression model with features X_M and parameters $\beta_M^* = n^{-1/2}\theta_M^*$. Therefore any statement about consistency and weak convergence that follows is a statement about this sequence of data generating mechanisms.

The event (17) can be rewritten as

$$\left\{\text{diag}(s_M)\left[\bar{\theta}_M - (n^{-1/2}\lambda)\mathcal{I}_M^{-1}s_M\right] + R_M \geq 0\right\}, \tag{19}$$

where

$$R_M = (n^{-1/2}\lambda) \cdot \left(\mathcal{I}_M^{-1} - (I_M(\hat{\beta}_M)/n)^{-1}\right) s_M$$

$$= (n^{-1/2}\lambda) \cdot \left(\mathcal{I}_M^{-1} - E_F[I_M(n^{-1/2}\theta_M^*)/n]^{-1}\right) s_M$$

$$+ (n^{-1/2}\lambda) \cdot \left(E_F[I_M(n^{-1/2}\theta_M^*)/n]^{-1} - (I_M(n^{-1/2}\theta_M^*)/n)^{-1}\right) s_M$$

$$+ (n^{-1/2}\lambda) \cdot D_M(n^{-1/2}\tilde{\theta}_M^*)(n^{-1/2}\theta_M^* - \hat{\beta}_M)s_M$$

is an unobservable remainder. The second equality is just Taylor's theorem with D_M denoting the derivative of I_M/n^{-1} with respect to β which is evaluated at some $n^{-1/2}\tilde{\theta}_M^*$ between $n^{-1/2}\theta_M^*$ and $\hat{\beta}_M$. If $\lambda = Cn^{1/2}$ then all terms in the above event have non-degenerate limits as $n \to \infty$ with the only randomness in the event being $\bar{\theta}_M$ and the remainder R_M.

Pre-selection, the first term of the remainder is seen to converge to 0 by the assumption that the information converges. The second term is seen to converge to 0 by the strong law of large numbers and the third term is seen to converge to 0 when $\hat{\beta}_M$ is consistent for $n^{-1/2}\theta_M^* =$ (both $\hat{\beta}_M^*$ and $\hat{\beta}_M^*$ converge to zero). As we are interested in the selective distribution we need to ensure that this remainder goes to 0 in probability, conditional on the selection event. For this it suffices to assume that the probability of selecting variables M is bounded below by some $\epsilon > 0$, ensuring that Lemma 1 of Tian & Taylor (2015) is applicable to transfer consistency pre-selection to consistency after selection. In terms of establishing a limiting distribution for inference we appeal to the CLT which holds for pre-selection and consider its behaviour after selection. It was shown in Tian & Taylor (2015) that CLTs that hold before selection extend to selective inference after adding noise to the outcome, under suitable assumptions.

We provide a sketch of such a proof in our setting. As we want to transfer a CLT pre-selection to the selective case we assume that $\bar{\theta}_M$ satisfies a CLT pre-selection under our sequence of data generating mechanisms. Now consider the selection event after removing the ignorable remainder R_M under the assumption that $\lambda = Cn^{1/2}$

$$\left\{ \operatorname{diag}(s_M) \left[\bar{\theta}_M - C \cdot \mathcal{I}_M^{-1} s_M \right] \geq 0 \right\}. \tag{20}$$

If the probability of (20) converges to some positive limit under our sequence of data generating mechanisms, it must agree with the same probability computed under the limiting Gaussian distribution. A direct application of the Portmanteau theorem establishes that the sequence of conditional distributions of $\bar{\theta}_M$ will therefore converge weakly and this limit will be the limiting Gaussian conditioned on (20).

This simple argument establishes weak convergence of the conditional distribution for a particular (θ_M^*, F) sending $n \to \infty$. For full inferential purposes, this pointwise weak convergence is not always sufficient. See Tibshirani et al. (2015) for some discussion of this topic and "honest" confidence intervals. A more rigorous treatment of transferring a CLT pre-selection to the selective model is discussed in Tian & Taylor (2015), where quantitative bounds are derived to compare the true distribution of a pivotal quantity to its distribution under the limiting Gaussian distribution.

Let us look at the inactive constraints. In the logistic regression example, with $\pi = \pi_M(\beta_M^*)$, we see that by construction

$$
\begin{aligned}
X_{-M}^\top W(z - X_M \hat{\beta}) &= X_{-M}^\top (y - \pi_M(\hat{\beta}_M)) \\
&= X_{-M}^\top (y - \pi_M(\beta_M^*)) - E_F[X_{-M}^\top W(\beta_M^*) X_M](\bar{\beta}_M - \beta_M^*) \\
&\quad + \Delta(F, M) s_M + R_{M,2},
\end{aligned}
$$

where

$$\Delta(F, M) = E_F[X_{-M}^\top W(\beta_M^*) X_M] E_F[(X_M^\top W(\beta_M^*) X_M)]^{-1}$$

is a population version of the matrix appearing in the well-known irrepresentable condition (Tropp, 2004; Wainwright, 2009) and the remainder $R_{M,2}$ goes to 0 in probability after appropriate rescaling Tian & Taylor (2015). Hence our "inactive constraints" can be rewritten in terms of the random

vector

$$X_{-M}^\top(y - \pi_M(\beta_M^*)) - E_F[X_{-M}^\top W(\beta_M^*)X_M](\bar\beta_M - \beta_M^*),$$

the remainder $R_{M,2}$ and a constant vector. Now, under our selected model, standard asymptotic arguments show that the random vector

$$\begin{pmatrix} \bar\beta_M - \beta_M^* \\ X_{-M}^\top(y - \pi) - E_F[X_{-M}^\top W(\beta_M^*)X_M](\bar\beta_M - \beta_M^*) \end{pmatrix} \tag{21}$$

satisifes a CLT before selection. It is straightforward to check that under this limiting Gaussian distribution these two components are independent. Indeed in the Gaussian case they are independent for every n. This implies a simplification similar to (5) occurs asymptotically for the problem (14). As this calculation was somewhat specific to logistic regression, this asymptotic independence of the two blocks and simplification of the constraints also holds when the likelihood in (14) is an exponential family and with β being the natural parameters.

If we knew β_M^* and F we could compute all relevant constants in the constraints and simply apply the polyhedral lemma (Lee et al., 2013). This would yield asymptotically exact selective inference for the selection event $\{(\hat M, s_{\hat M}) = (M, s_M)\}$ by construction of a pivotal quantity

$$g(\bar\theta_M; \lim_{n\to\infty} n E_F[I_M(n^{1/2}\theta_M^*)]^{-1}; A, b) \tag{22}$$

as in (6) , where A and b can be derived from the polyhedral constraints (19). Specifically $A = -\text{diag}(s_M)$, and $b = -E_F[I_M(n^{1/2}\theta_M^*)/n]^{-1}(n^{-1/2}\lambda)s_M) - n^{1/2}R_{M,1}$.

However the quantities needed to compute $\lim_{n\to\infty} n E_F[I_M(n^{1/2}\theta_M^*)]^{-1}$ are unknown, though there are certainly natural plug-in estimators that would be consistent without selection. This suggests using a plug-in estimate of variance. In Tian & Taylor (2015) it was shown that, under mild regularity assumptions, consistent estimates of variance can be plugged into limiting Gaussian approximations for asymptotically valid selective inference. Hence, to construct a practical algorithm, we apply the polyhedral lemma to the limiting distribution of $n^{1/2}\bar\beta_M$, with M, s_M fixed and $n\nabla^2\ell_M(\hat\beta_M)$ as a plugin estimate for $nE_F[I_M(\beta_M^*)]$.

Thus we have the following result:

Result 1. *Suppose that the model described in Section (3.1) holds for all n and some (θ_M^*, F) such that the corresponding population covariance $\mathcal{I}_M(\theta_M^*)$ is non-degenerate. Then the pivot (22) is asymptotically $U(0, 1)$ as $n \to \infty$ conditioned on having selected variables M with signs s_M. Furthermore plugging in $nI_M(\hat\beta_M)^{-1}$ both in the limiting variance and in the constraints of (19) of the pivot is also asymptotically $U(0, 1)$.*

As noted in the introduction a detailed proof of this result will appear elsewhere.

Remark A. *In the Gaussian model, our one-step estimator has the form*

$$\bar\beta_M = \hat\beta_M + (X_M^\top X_M)^{-1}X_M^\top(y - X\hat\beta_M) = (X_M^\top X_M)^{-1}X_M^\top y \tag{23}$$

with $\bar\beta_M \sim N((X_M^\top X_M)^{-1}X_M^\top X\beta, (X_M^\top X_M)^{-1}\sigma^2)$ and constraints

$$\left\{\text{diag}(s_M)\left[\bar\beta_M - (X_M^\top X_M)^{-1})^{-1}\lambda s_M)\right] \geq 0\right\}. \tag{24}$$

These are just the usual least squares estimates for the active variables. We note the strong similarity between the one-step estimator and the "debiased lasso" construction of Zhang & Zhang (2014), Bühlmann (2013), van de Geer et al. (2013), and Javanmard & Montanari (2014). In the context of Gaussian regression, the latter approach uses

$$\hat{\beta}^d \equiv \hat{\beta}_M + (1/N)\Theta X^\top (y - X\hat{\beta}_M) \tag{25}$$

where Θ is an estimate of $(X^\top X/N)^{-1}$. This estimator takes a Newton step in the full model direction. Our one-step estimator has a similar form to (25), but takes a step only in the active variables, leaving the others at 0. Further the $(X_M^\top X_M/N)^{-1}$ is used as the estimate for Θ. The debiased lasso (25) uses a full model regularized estimate of Θ and ignores the constraints in (17). As pointed out by a referee the debiased lasso is more complex because it does not assume that the lasso has the screening property, (i.e., the true nonzero set is included in the estimated nonzero set). Another important difference is that their target of inference for the debiased lasso is a population parameter, that is, is determined before observing the data. This is not the case for our procedure.

Remark B.   *The conclusions of Result 1 can be strengthened to hold uniformly over compact subsets of θ_M^* parameters. As we do not pursue this here, such results are stated more formally in Tian and Taylor (2015) in the setting where noise is first added to the data before model fitting and selection. Lemma 1 is a statement about the conditional distribution of the pivot under selected model. In the Gaussian case similar results hold unconditionally for the pivot in the saturated model Tian & Taylor (2014); Tibshirani et al. (2015).*

Remark C.   *For the Cox model we define the one-step estimator in a similar fashion. In terms of the appropriate distribution for inference we simply replace the likelihood by a partial likelihood.*

3.3. Unpenalized Variables

It is common to include an intercept in logistic regression and other models, which is typically not penalized in the ℓ_1 penalty. More generally suppose that features $U_{n \times k}$ are to have unpenalized coefficients, whereas those for $X_{n \times p}$ are to be penalized. This changes the KKT conditions we have been using somewhat, but not in any material way. We now have $\eta = U\alpha + X\beta$ and the KKT conditions now include a set of conditions for the unpenalized variables, say U. In the logistic regression case these read as

$$U^\top W(z - X_M\hat{\beta}_M - U\hat{\alpha}) = 0. \tag{26}$$

The corresponding one-step estimator is

$$\begin{pmatrix} \bar{\alpha} \\ \bar{\beta}_M \end{pmatrix} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_M \end{pmatrix} + I_M(\hat{\alpha}, \hat{\beta}_M)^{-1} \begin{pmatrix} 0 \\ \lambda s_M \end{pmatrix}$$

where I_M is the $(|M| + k) \times (|M| + k)$ Fisher information matrix of the submodel M. In terms of constraints we only really need consider the signs of the selected variables and the corresponding "plug-in" form of the active constraints are

$$\text{sign}\left(\bar{\beta}_M - E_M \left(I_M(\hat{\alpha}, \hat{\beta}_M)^{-1} \begin{pmatrix} 0 \\ \lambda s_M \end{pmatrix} \right) \right) = s_M.$$

The population version uses the expected Fisher information at (α^*, β_M^*) instead of the observed information and E_M is the matrix that selects rows corresponding to M. As our one-step estimator

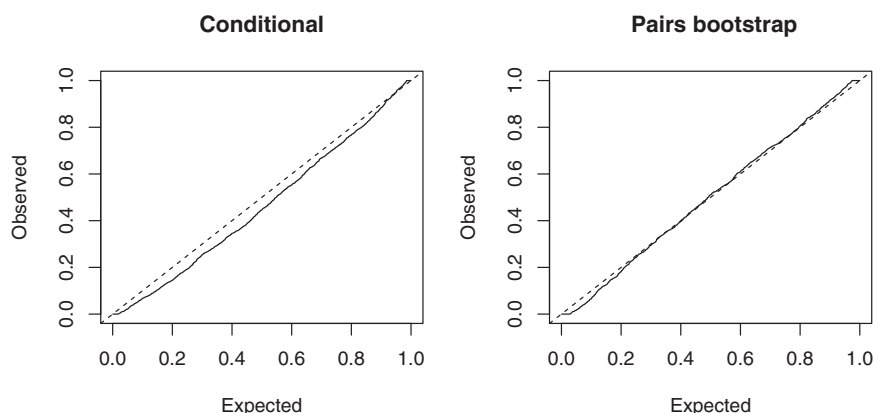**Conditional**   **Pairs bootstrap**

FIGURE 2: *P*-values for the lasso in the Gaussian setting: $n = 200$, $P = 50$, 20 strong signals. Predictors have pairwise correlation of 0.3 and variance of the errors depends on non-signal variables. Shown are the quantile–quantile plots for the non-signal variables in realizations for which the lasso has successfully screened (captured all of the signal variables). We see that the conditional analysis yields anti-conservative *P*-values, whereas the pairs bootstrap gives *P*-values closer to uniform.

is expressed in terms of the likelihood this estimator can be used in problems that are not regression problems but that have unpenalized parameters such as the graphical LASSO discussed in Section 6.

3.4. The Random *X*-Case

The truncated Gaussian theory of Lee et al. (2013) assumes that X is fixed, and conditions on it in the inference. When X is random (most often the case), this ignores its inherent variability and makes the inference non-robust when the error variance is non-heterogeneous. This point is made forcefully by Buja et al. (2016).

The one-step estimation framework of this article provides a way to deal with the problem. Consider for simplicity the Gaussian case for the lasso of Lee et al. (2013), which expresses the selection as $Ay \leq 0$ with $y \sim N(\mu, I\sigma)$. Above we have re-expressed this as $\bar{\beta}_M \sim N(\beta^*, \Sigma)$ where $\bar{\beta}_M$ is the one-step estimator for the selected model. In the Gaussian case, $\bar{\beta}_M$ is just $\hat{\beta}_M$, the usual least squares estimate on the selected set and $\Sigma = (X_M^\top X_M)^{-1}\sigma^2$. Now analogous to (21), for the Gaussian case we have the asymptotic result

$$\begin{pmatrix} \bar{\beta}_M - \beta_M^* \\ X_{-M}^\top(y - X\beta^*) - E_F[X_{-M}^\top X_M](\bar{\beta}_M - \beta_M^*) \end{pmatrix} \sim N(0, \Sigma'). \qquad (27)$$

This suggests that we can use the pairs bootstrap to estimate the *unconditional* variance–covariance matrix Σ' and then simply apply the polyhedral lemma, as before. Alternatively a sandwich-style estimator of Σ' can be used.

Figure 2 shows an example illustrating how the pairs bootstrap can give robustness with heterogeneity of the error variance. Details are in the caption.

4. SIMULATIONS

To assess performance in the ℓ_1-penalized logistic model we generated Gaussian features with pairwise correlation 0.2 in two scenarios: $n = 30$, $P = 10$ and $n = 40$, $P = 60$. Then y was generated as $\Pr(Y = 1|x) = 1/(1 + \exp(-x^\top \beta))$. There are two signal settings: null ($\beta = 0$) (Figure 3)
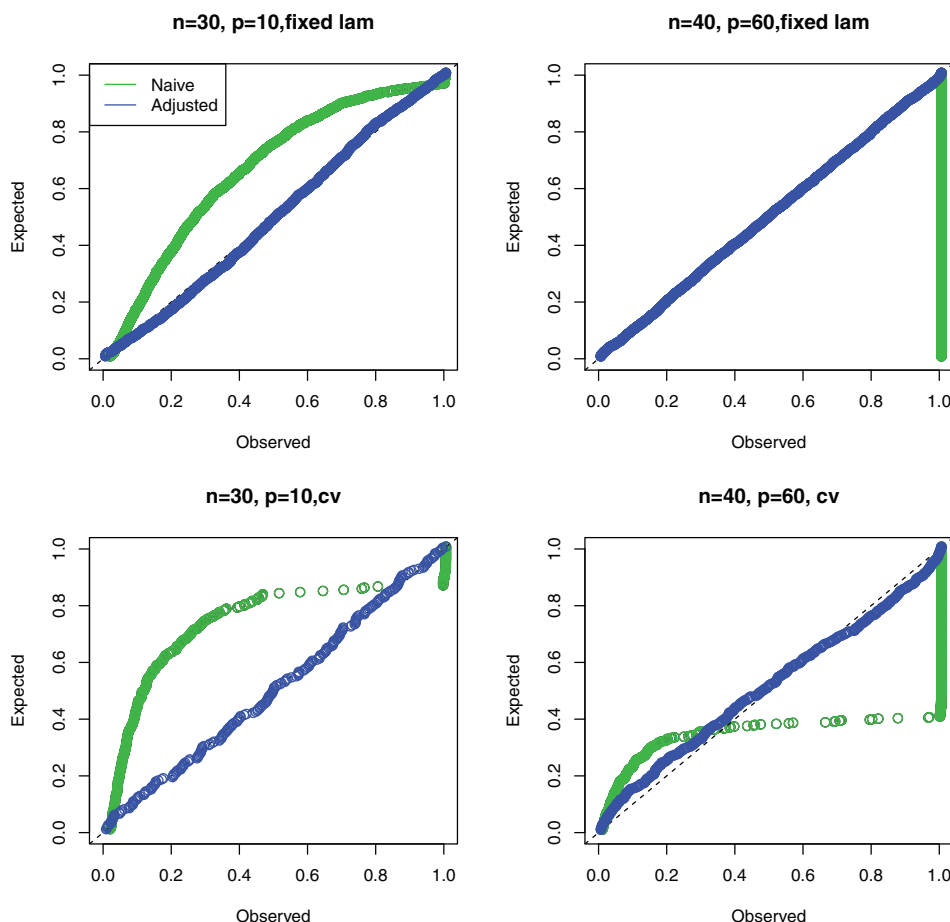
FIGURE 3:   *P*-values for the logistic regression model, null setting. The top panels use a fixed λ, whereas the bottom ones use cross-validation to choose λ.

non-null ($\beta = (5, 0, 0 \ldots)$) (Figure 4). Finally in each case we tried two methods for choosing the regularization parameter λ: a fixed value yielding a moderately sparse model and cross-validation. The figures show the cumulative distribution function of the resulting *P*-values over 1,000 simulations. Thus a function above the 45 degree line indicates an anti-conservative test in the null setting and a test with some power in the non-null case. We see the adjusted *P*-values are close to uniform under the null in every case and show power in the non-null setting. Even with a cross-validation-based choice for λ the type I error seems to be controlled, although we have no theoretical support for this finding. In Figure 3 we also plot the naive *P*-values from GLM theory: as expected they are very anti-conservative.

Figure 5 shows the results of an analogous experiment for the Cox model in the null setting, using exponential survival times and random 50% censoring. Type I error control is good, except in the cross-validation case where it is badly anti-conservative for smaller *P*-values. We have seen similar behaviour in the Gaussian lasso setting, and this phenomenon deserves further study.

Table 1 shows the miscoverage and median lengths of intervals for the logistic regression example in the null setting, with a target miscoverage of 10%. The intervals can sometimes be very long, and in fact, have infinite expected length.
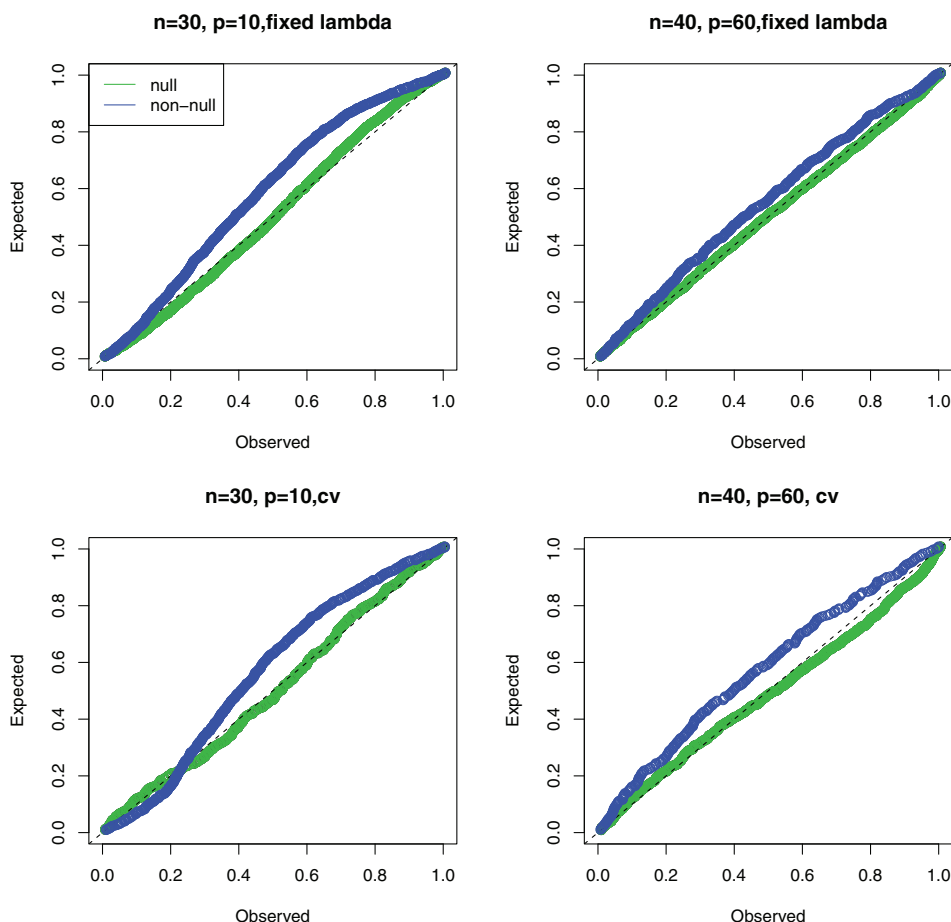
FIGURE 4: *P*-values for the logistic regression model, non-null setting. The top panels use a fixed λ, whereas the bottom ones use cross-validation to choose λ.

As a comparison Table 2 shows analogous results for Gaussian regression, using the proposal of Lee et al. (2013). For estimation of the error variance σ^2 we used the mean residual error for $N > p$ and the cross-validation estimate of Reid, Tibshirani, & Friedman (2013) for $N < p$.

Again the intervals can be quite long. There are potentially better ways to construct the intervals: Tibshirani et al. (2015) propose a bootstrap method for post-selection inference that in our current problem would draw bootstrap samples from $z_1, z_2, \ldots z_n$ and use their empirical distribution in the polyhedral lemma (in place of the Gaussian distribution). The "randomized response" strategy provides another way to obtain shorter intervals, at the expense of increased computation: see Tian & Taylor (2015).

5. EXAMPLES

5.1. Liver Data

The data in this example and the following (edited) description were provided by D. Harrington and T. Fleming.

"Primary biliary cirrhosis (PBC) of the liver is a rare but fatal chronic liver disease of unknown cause, with a prevalence of about 50-cases-per-million population. The primary pathologic event
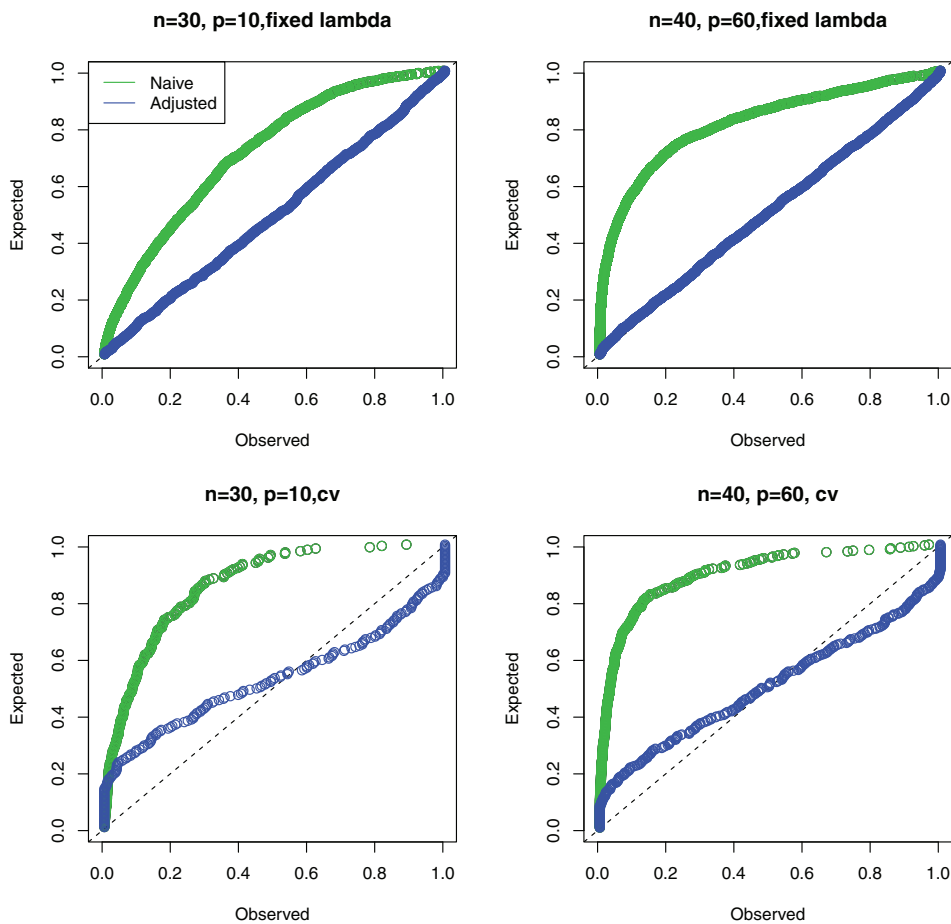
FIGURE 5: *P*-values for the Cox Model, null setting. The top panels use a fixed λ, whereas the bottom ones use cross-validation to choose λ.

TABLE 1: Lasso-penalized logistic regression.

	$N > P$, fixed λ	$N > P$, cv	$N < P$, fixed λ	$N < P$, cv
Miscoverage	0.11	0.11	0.14	0.14
Median length	8.07	6.63	5.75	29.69

Note: Details as in Figure 4. Shown are the miscoverage and median length of selection (confidence) intervals for main proposal of this article. The target miscoverage is 10%. Selection of λ is done using either a fixed value yielding moderate sparsity or cross-validation (cv).

appears to be the destruction of interlobular bile ducts, which may be mediated by immunologic mechanisms.

The following briefly describes data collected for the Mayo Clinic trial in PBC of the liver conducted between January, 1974 and May, 1984 comparing the drug D-penicillamine (DPCA) with a placebo. The first 312 cases participated in the randomized trial of D-penicillamine versus placebo, and contain largely complete data. An additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival.

TABLE 2: Lasso-penalized Gaussian regression.

	$N > p$, fixed λ	$N > p$, cv	$N < p$, fixed λ	$N < p$, cv
Miscoverage	0.11	0.12	0.13	0.10
Median length	12.29	6.90	20.50	32.06

Note: Shown are the miscoverage and median lengths of selection (confidence) intervals based on the Gaussian model of Lee et al. (2013). The target level is 10%. Selection of λ is done using either a fixed value yielding moderate sparsity or cross-validation (cv).

Six of those cases were lost to follow-up shortly after diagnosis, so there are data here on an additional 106 cases as well as the 312 randomized participants."

We discarded observations with missing values, leaving 276 observations. The predictors are

X_1 : Treatment Code, 1 = D-penicillamine, 2 = placebo.

X_2 : Age in years. For the first 312 cases age was calculated by dividing the number of days between birth and study registration by 365.

X_3 : Sex, 0 = male, 1 = female.

X_4 : Presence of ascites, 0 = no, 1 = yes.

X_5 : Presence of hepatomegaly, 0 = no, 1 = yes.

X_6 : Presence of spiders, 0 = no, 1 = yes.

X_7 : Presence of edema, 0 = no, 0.5 yes but responded to diuretic treatment, 1 = yes, did not respond to treatment.

X_8 : Serum bilirubin, in mg/dl.

X_9 : Serum cholesterol, in mg/dl.

X_{10} : Albumin, in gm/dl.

X_{11} : Urine copper, in μg/day.

X_{12} : Alkaline phosphatase, in U/l.

X_{13} : SGOT, in U/ml.

X_{14} : Triglycerides, in mg/dl.

X_{15} : Platelet count; coded value is number of platelets per cubic ml. of blood divided by 1,000.

X_{16} : Prothrombine time, in seconds.

X_{17} : Histologic stage of disease, graded 1, 2, 3, or 4.

We applied Cox's proportional hazards model. Figures 6 and 7 show the results. As expected the adjusted P-values are larger than the naive ones and the corresponding selection (confidence) intervals tend to be wider.
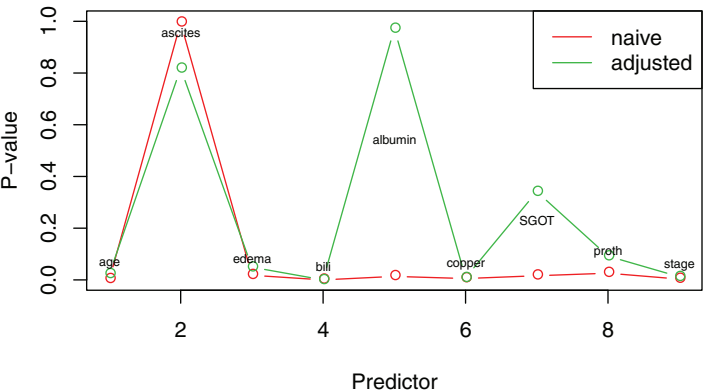

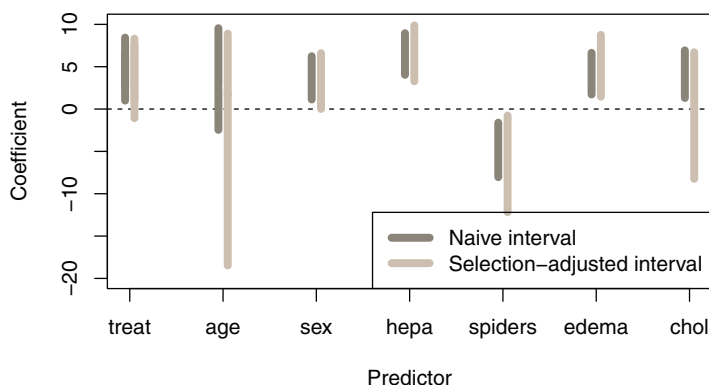FIGURE 6: *P*-values for Cox model applied to the liver data.

FIGURE 7: Selection intervals for Cox model applied to the liver data.

6. THE GRAPHICAL LASSO

Another, different, example is the graphical lasso for estimation of sparse inverse covariance graphs. Here we have data $X_{n \times p} \sim N(0, \Sigma)$. Let $S = X^\top X/N$, $\Theta = \Sigma^{-1}$.

We maximize

$$\ell(\Theta) = \log \det \Theta - \operatorname{tr} S\Sigma - \lambda ||\Theta||_1, \tag{28}$$

where the norm in the second is the sum of the absolute values.

The KKT conditions have the form

$$\Theta^{-1} - S - \lambda s(\Theta) = 0 \tag{29}$$

or for one row/column, $\Sigma_{11}\beta - s_{12} - \lambda s(\beta) = 0$, where β is a $p - 1$ vector in the pth row/column of Θ, excluding the diagonal, and Σ_{11} is the block of Σ excluding one row and column. Defining $R = d^2\ell/d\Theta d\Theta^\top$ we have

$$\bar{\Theta}_M = \hat{\Theta}_M + \lambda R^{-1} s_M. \tag{30}$$

Hence we apply the polyhedral lemma to $\bar{\Theta}_M \sim N(\Theta_M^*, R^{-1})$ with constraints $-\operatorname{diag}(s_M)(\bar{\Theta}_M - R^{-1}s_M) \leq 0$. From this we can obtain P-values for testing whether a link parameter is zero ($H_0 : \theta_{jk} = 0$) and confidence intervals for θ_{jk}. We note the related work on high-dimensional inverse covariance estimation in Jankova & van de Geer (2014).

Figure 8 shows the results of a simulation study with $n = 80$, $P = 20$ with the components $X_1, X_2, \ldots X_p$ being standard Gaussian variables. All components were generated independently except for the first two, which had correlation 0.7. A fixed moderate value of the regularization parameter was used. Conditioning on realizations for which the partial correlation for the first two variables was non-zero, the Figure shows the P-values for non-null (1,2) entry and the null (the rest). We see that the null P-values are close to uniform and the non-null ones are (slightly) sub-uniform.

6.1. Example

Here we analyze the protein data discussed in Friedman, Hastie, & Tibshirani (2008). The measurements are from flow cytometry, with 11 proteins measured over 7,466 cells. Table 3 and Figure 9 show the results of applying the post-selection procedure with a moderate value of the
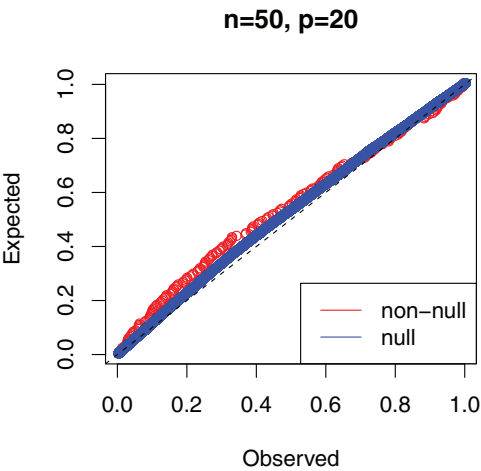
n=50, p=20

FIGURE 8:   Simulation results: *P*-values for the graphical lasso. Details are given in the text.

TABLE 3: *P*-values from the graphical lasso applied to the protein data.

Protein pair	*P*-values
Raf -Mek	0.789
Mek -P38	0.005
Plcg- PIP2	0.107
PIP2 -P38	0.070
PKA -P38	0.951
P38 -Jnk	0.557
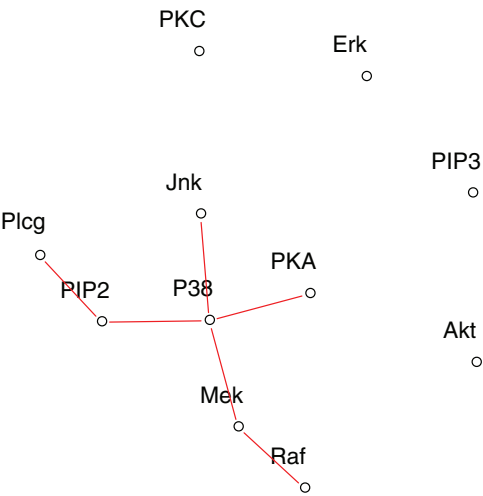
FIGURE 9:   Results for protein data. The red lines indicate non-zero fitted entries in $\hat{\Theta}$.

regularization parameter λ. Six interactions are present in the selected model, with only one (Mek-P38) being strongly significant.

7. DISCUSSION

We have proposed a method for post-selection inference, with applications to ℓ_1-penalized likelihood models. These include generalized linear models, Cox's proportional hazards model, and the graphical lasso. As noted earlier as our focus has been on selection via ℓ_1-penalization, a similar approach can be applied to forward stepwise methods for likelihood models, and in principle, least angle regression.

A major challenge remains in the estimation of the tuning parameter λ. One possibility is to use a choice proposed by Negahban et al. (2012) (Theorem 1), which is $\lambda' = 2 \cdot \mathrm{E}[\max_j \nabla L(0)]$. In the logistic and Cox models, for example, this is not a function of y and hence the proposals of this article can be applied. More generally it would be desirable to allow for the choice of λ by cross-validation in our methodology. Choosing λ by cross-validation is feasible, particularly in the "randomized response" setting of Tian & Taylor (2015) though this approach requires MCMC for inference. A related approach to selective inference after cross-validation is described in Loftus (2015).

The proposals of this article are implemented in our `selectiveInference` R package in the public CRAN repository.

ACKNOWLEDGEMENTS

APPENDIX

The Polyhedral Lemma and Truncation Limits for Post-Selection Gaussian Inference

Let $y \sim N(\mu, \Sigma)$ and suppose that we apply the lasso with parameter λ to data (X, y), yielding active variables M. Lee et al. (2013) show that for the active variables, $A_1 = -D(X_M^\top X_M)^{-1} X_M^\top$, $b_1 = -D(X_M^\top X_M)^{-1} \lambda s$ where $D = \mathrm{diag}(s)$. For inactive variables $A_0 = \frac{1}{\lambda} \begin{pmatrix} X_{-M}^\top \\ -X_{-M}^\top \end{pmatrix}$, $b_0 = \begin{pmatrix} 1 + X_{-M}^\top X_M \hat{\beta}/\lambda \\ 1 - X_{-M}^\top X_M \hat{\beta}/\lambda \end{pmatrix}$.

We finally define $A = \begin{pmatrix} A_1 \\ A_0 \end{pmatrix}$, $b = (b_1, b_0)$. They also show that

$$\{Ay \le b\} = \{\mathcal{V}^-(y) \le \gamma^\top y \le \mathcal{V}^+(y), \ \mathcal{V}^0(y) \ge 0\}; \tag{31}$$

and furthermore, $\gamma^\top y$ and $(\mathcal{V}^-(y), \mathcal{V}^+(y), \mathcal{V}^0(y))$ are statistically independent. This surprising result is known as the *polyhedral lemma*. Let $c \equiv \Sigma\gamma(\gamma^\top \Sigma \gamma)^{-1}$, and $r \equiv (I_N - c\gamma^\top)y$. Then the three values on the righthand side of (31) are computed via

$$\mathcal{V}^-(r) = \max_{j:(Ac)_j<0} \frac{b_j - (Ar)_j}{(Ac)_j}$$

$$\mathcal{V}^+(r) = \min_{j:(Ac)_j>0} \frac{b_j - (Ar)_j}{(Ac)_j}$$

$$\mathcal{V}^0(r) = \min_{j:(Ac)_j=0} b_j - (Ar)_j. \tag{32}$$

Hence the selection event $\{Ay \leq b\}$ is equivalent to the event that $\gamma^{\top} y$ falls into a certain range, a range depending on A and b. This equivalence and the independence means that the conditional inference on $\gamma^{\top} \mu$ can be made using the truncated distribution of $\gamma^{\top} y$, a truncated normal distribution.

The Hessian for the Graphical Lasso

Let $(\Delta_{ij})_{1 \leq i \leq j \leq p}$ denote the upper triangular parameters for the graphical lasso and

$$\Theta(\Delta)_{ij} = \begin{cases} \Delta_{ji} & i > j \\ \Delta_{ij} & i \leq j \end{cases}$$

be the symmetric version so that

$$\frac{\partial}{\partial \Delta_{ij}} \Theta_{kl} = \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}.$$

Now,

$$\frac{\partial^2}{\partial \Theta_{ij} \partial \Theta_{kl}}(-\log \det \Theta) = \text{Tr}(e_i e_j^{\top} \Theta^{-1} e_k e_l^{\top} \Theta^{-1}) = \Sigma_{jk} \Sigma_{il}$$

with $\Sigma = \Theta^{-1}$. Note that we evaluate this at a symmetric matrix, that is, $\Theta^{\top} = \Theta$.

Therefore,

$$\frac{\partial^2}{\partial \Delta_{ij} \partial \Delta_{kl}}(-\log \det(\Theta(\Delta))) = \sum_{i',j',k',l'} \Sigma_{j'k'} \Sigma_{i'l'}(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{ji'})(\delta_{kk'}\delta_{ll'} + \delta_{kl'}\delta_{lk'})$$

$$= 2(\Sigma_{il} \Sigma_{jk} + \Sigma_{ik} \Sigma_{jl}).$$

BIBLIOGRAPHY

Bickel, P. J., Klaassen, C. A., Ritov, Y., & Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins Univ. Press, Baltimore.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models, *Bernoulli*, 19(4), 1212–1242.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhang, K., & Zhao, L. (2016). *A Conspiracy of Random X and Nonlinearity Against Classical Inference in Linear Regression*. submitted.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l1 and l1+l2 penalization. *Electronic Journal of Statistics*, 2, 1153–1194. URL: http://dx.doi.org/10.1214/08-EJS287

Fithian, W., Sun, D., & Taylor, J. (2014). Optimal inference after model selection. *ArXiv e-prints*.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9, 432–441.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall, London.

Jankova, J. & van de Geer, S. (2014). Confidence intervals for high-dimensional inverse covariance estimation. *arXiv:1403.6752*.

Javanmard, A. & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15, 2869–2909.

Lee, J., Sun, D., Sun, Y., & Taylor, J. (2013). Exact post-selection inference, with application to the Lasso. *arXiv:1311.6238*.

Loftus, J. R. (2015). Selective inference after cross-validation. *ArXiv e-prints*.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., & Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4), 538–557. URL: http://dx.doi.org/10.1214/12-STS400

Reid, S., Tibshirani, R., & Friedman, J. (2013). A study of error variance estimation in lasso regression. *ArXiv e-prints* to appear in *Statistica Sinica*.

Taylor, J., Lockhart, R., Tibshirani, R., & Tibshirani, R. (2014). Post-selection adaptive inference for least angle regression and the Lasso. *arXiv: 1401.3889*; submitted.

Taylor, J. & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25), 7629–7634. URL: http://www.pnas.org/content/112/25/7629.abstract

Tian, X. & Taylor, J. E. (2014). Asymptotics of selective inference. *ArXiv e-prints*.

Tian, X. & Taylor, J. E. (2015). Selective inference with a randomized response. *ArXiv e-prints*.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., & Wasserman, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. *ArXiv e-prints*.

Tibshirani, R. (2013). The Lasso problem and uniqueness. *Electronic Journal of Statistics*, 7, 1456–1490.

Tropp, J. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10), 2231–2242.

van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv: 1303.0518v2*.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55, 2183–2202.

Zhang, C.-H. & Zhang, S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *Journal of the Royal Statistical Society Series B*, 76, 217–242.