Selective inference after cross-validation

Joshua Loftus

e-mail: joftius@stanford.edu

Department of Statistics
Sequoia Hall
390 Serra Mall
Stanford, CA

Abstract: This paper describes a method for performing inference on models chosen by cross-validation. When the test error being minimized in cross-validation is a residual sum of squares it can be written as a quadratic form. This allows us to apply the inference framework in Loftus et al. (2015) for models determined by quadratic constraints to the model that minimizes CV test error. Our only requirement on the model training procedure is that its selection events are regions satisfying linear or quadratic constraints. This includes both Lasso and forward stepwise, which serve as our main examples throughout. We do not require knowledge of the error variance σ^2. The procedures described here are computationally intensive methods of selecting models adaptively and performing inference for the selected model. Implementations are available in an R package.

1. Introduction

We consider a modeling scenario with outcome or response variable $y \sim N(\mu, \sigma^2 I)$ and a matrix X of predictor variables. We hope to model the response as $y = X\beta + \epsilon$ for some unknown β, and imagine that β is sparse in the sense that few of its entries are nonzero. In this setting various model selection procedures exist that select a subset of predictors X_A with the hope that $X_A\beta_A$ is a good approximation of μ. For example, forward stepwise begins with an empty model and then sequentially adds the most predictive variable at each step. After k steps the forward stepwise model includes k predictors. The Lasso estimator is defined as

$$\hat{\beta}_\lambda := \arg\min_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{1}$$

and has the property that its solutions are increasingly sparse as λ increases. For large enough λ, $\beta_\lambda = 0$, and if λ decreases continuously then predictors will enter the Lasso model one at a time in a similar fashion to forward stepwise. In this way, both forward stepwise and Lasso algorithms can output models in a sequence of increasing model complexity, i.e. number of estimated parameters.

For such model selection procedures one of the most important questions in their practical application is *how complex of a model should we use?* How many steps should we allow forward stepwise to take? What value of λ will yield a reasonable Lasso model? There are many answers to these questions motivated

by theory, but in practice cross-validation is perhaps the most widely used. In K-fold cross-validation, the data are partitioned into K independent subsets $D_1, \ldots, D_K$.. For each $k = 1, \ldots, K$, subset D_k is held out while the rest of the data are used to train a sequence of predictive models $\hat{f}^k_\lambda$. The subscript λ indexes model complexity, and might take continuous values for the Lasso, or integer values for forward stepwise. The predicted value $\hat{y}^k_\lambda = \hat{f}^k_\lambda(x^k)$, and with a predictive loss function $\ell(y, \hat{y})$ the cross-validation estimate of prediction error is given by

$$\text{cv-Error}_\ell(\lambda) := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|D_k|} \ell(y^k, \hat{y}^k_\lambda) \tag{2}$$

Throughout this paper we use squared-error loss $\ell(y, \hat{y}) = \|y - \hat{y}\|_2^2$ and suppress the ℓ notation unless otherwise specified. Finally, to choose λ it is reasonable to pick the minimizer $\hat{\lambda}$ of (2) with the hope that a model with complexity $\hat{\lambda}$ will have low prediction error. To simplify notation we also assume K divides n and all folds have equal size $|D_k| = n/K$, and consider the equivalent problem of minimizing

$$\text{cv-Error}(\lambda) = \sum_{k=1}^{K} \|y^k - \hat{y}^k_\lambda\|_2^2. \tag{3}$$

Much recent work on selective inference shows how to perform inference for Lasso and forward stepwise when the model complexity λ has been specified independently of the data Lee et al. (2015); Taylor, Loftus and Tibshirani (2015); Tibshirani et al. (2014). Principled choices of λ can be made if σ^2 is known. When σ^2 is unknown, Tian, Loftus and Taylor (2015) use the square-root Lasso of Belloni, Chernozhukov and Wang (2010), and Gross, Taylor and Tibshirani (2015); Loftus and Taylor (2015) show how to do selective F tests. Cross-validation remains one of the most popular methods in practice, and no previous work has shown how to conduct inference conditional on models selected this way. We do this now, leveraging the quadratic model selection framework described in Loftus and Taylor (2015).

1.1. Selective hypothesis tests

Loftus and Taylor (2015) describe a framework for significance tests adjusted for model selection when the choice of selected model is determined by quadratic constraints. That is, denoting the selected model as $M(y)$, we have for each possible model m

$$M(y) = m \iff y \in \bigcap_{j \in J_m} \{y : y^T Q_{m,j} y + a_{m,j}^T y + b_{m,j} \geq 0\} \tag{4}$$

for some finite index set J_m, and $Q_{m,j}, a_{m,j}, b_{m,j}$ depend on y only through m. This definition is general enough to include the Lasso, marginal screening, forward stepwise, and forward stepwise with groups of variables. In the present

work we show that cross-validation combined with any of the above also fit within this framework.

Given a model m there is an active set A_m of variables included in the model. For each $j \in A_m$ we wish to conduct a significance test. Assuming

$$y \sim N(\mu, \sigma^2 I) \tag{5}$$

we give a test for each of the parameters $\beta_{A_m,j}$ with $j \in A_m$ where β_{A_m} is the population least squares parameter vector $P_{A_m}\mu$, the projection of μ onto the column space of X_{A_m}. If the linear model is correctly specified and A_m contains the true active set, then the null hypotheses for variables which are null in the true model also hold in the selected model. We use the classical t, χ^2, or F significance tests for single coordinates or groups of coordinates of β_{A_m}, and compute p-values from the null distributions truncated to the model selection region. For a variable $j \in A_m$, the selective null hypothesis and type 1 error are

$$\begin{aligned} &H_0(A_m, j) : \beta_{A_m,j} = 0 \\ &\mathbb{P}_{m,H_0(A_m,j)}(\text{reject } H_0(A_m, j) | M(y) = m) \end{aligned} \tag{6}$$

where the probability under m and $H_0(A_m, j)$ is computed from the probability model (5) truncated to the region implied by $M(y) = m$ and with the constraint on μ implied by $H_0(A_m, j)$. Further details on χ^2 and F tests are provided in Loftus and Taylor (2015), and Lee et al. (2015) discuss z-tests in the setting where model selection is affine rather than quadratic. Fithian, Sun and Taylor (2014) discuss the advantages of selective inference and optimality theory in exponential families.

2. Test error estimates as quadratic forms

This section shows how we can write test error estimates as quadratic forms. Thus, when a model is chosen by minimizing these estimates we can apply the strategy in Loftus and Taylor (2015) to conduct inference conditional on the chosen model. For simplicity we first illustrate the approach with forward stepwise, writing s to index steps rather than λ, and then discuss how the Lasso and other methods fall within the same framework.

2.1. A single training-test split

Before analyzing full cross-validation, we first consider splitting the data into independent partitions $(y^{\text{tr}}, X^{\text{tr}})$ for training and $(y^{\text{te}}, X^{\text{te}})$ for estimating test error. For concreteness assume we run forward stepwise for a fixed number of steps S and observe

$$M^{\text{tr}}(y^{\text{tr}}) = m \iff y^{\text{tr}} \in \bigcap_{j \in J} E^{\text{tr}}_{m,j} \tag{7}$$

where $E^{\text{tr}}_{m,j}$ is an event of the form $\{z \in \mathbb{R}^{|\text{tr}|} : z^T Q^{\text{tr}}_{m,j} z + A^{\text{tr}}_{m,j} z + b^{\text{tr}}_{m,j} \geq 0\}$ and $|\text{tr}|$ is the size of the training set. For each $s = 1, \ldots, S$ we have a model in the forward stepwise path with an associated parameter vector $\hat{\beta}_{m,s}$ given by

$$\hat{\beta}_{m,s} = (X^{\text{tr}}_{m,s})^{\dagger} y^{\text{tr}} \tag{8}$$

Picking the model in this forward stepwise path with the smallest test error means finding

$$\hat{s} := \arg\min_s \text{RSS}^{\text{te}}_{m,s}(y^{\text{te}}) = \arg\min_s \|y^{\text{te}} - X^{\text{te}}_{m,s}\hat{\beta}_{m,s}\|_2^2 \tag{9}$$

Denote $P_{m,s} := X^{\text{te}}_{m,s}(X^{\text{tr}}_{m,s})^{\dagger}$, so the RSS criterion above is $\|y^{\text{te}} - P_{m,s} y^{\text{tr}}\|_2^2$. The following Lemma follows from simple algebra and the definitions.

Lemma 2.1. *For all $r \neq s$ with $1 \leq r \leq S$, define*

$$E^{te}_{m,r} := \{y : \begin{bmatrix} y^{tr} \\ y^{te} \end{bmatrix}^T \begin{bmatrix} P^T_{m,s}P_{m,s} - P^T_{m,r}P_{m,r} & -(P_{m,s} - P_{m,r})^T \\ -(P_{m,s} - P_{m,r}) & 0 \end{bmatrix} \begin{bmatrix} y^{tr} \\ y^{te} \end{bmatrix} \leq 0\} \tag{10}$$

Then conditional on the models fitted on the training set,

$$\hat{s} = s \iff y \in \bigcap_{\substack{r=1 \\ r \neq s}}^{S} E^{te}_{m,r}. \tag{11}$$

Next we combine all the observed inequalities. We can pad the matrices defining each $E^{\text{tr}}_{m,j}$ by adding a block of zeroes for y^{te}, so that

$$y^{\text{tr}} \in E^{\text{tr}}_{m,j} \iff y \in E_{m,j}$$

We have established

Proposition 2.1. *The model selection event determined by a single training and test split decomposes as the following intersection of quadratic inequalities*

$$M^{tr}(y^{tr}) = m \text{ and } \hat{s} = s \iff y \in \bigcap_{j \in J} E_{m,j} \cap \bigcap_{r=1}^{S} E^{te}_{m,j} \tag{12}$$

where J is the (finite) index set given in (7).

2.2. K-fold cross-validation

To extend the argument to cross-validation we first need more notation. Let $1 \leq f \leq K$ index CV folds. Write X^f for the test set and X^{-f} for the training set for fold f. The model m_f is trained on (y^{-f}, X^{-f}), and there is an associated event

$$M^f(y^{-f}) = m_f \iff y^{-f} \in \bigcap_{j \in J_f} E^f_{m,j}. \tag{13}$$

For each sparsity s there is a corresponding fit $\hat{\beta}_{m_f,s}$. Suppose we use least squares on the active set, so $\hat{\beta}_{m_f,s} = (X^{-f}_{m_f,s})^\dagger y^{-f}$. We choose s minimizing the cv-RSS,

$$s = \arg\min_s \sum_{f=1}^{K} \|y^f - X^f_{m_f,s}\hat{\beta}_{m_f,s}\|_2^2 \tag{14}$$

Define

$$P_{f,s} := X^f_{m_f,s}(X^{-f}_{m_f,s})^\dagger \tag{15}$$

The objective in (14) can be written

$$\begin{aligned}
\mathrm{RSS}(s) &:= \sum_{f=1}^{K} \|y^f - P_{f,s}y^{-f}\|_2^2 \\
&= \sum_{f=1}^{K} \left(\|y^f\|_2^2 - (y^f)^T P_{f,s}y^{-f} - (y^{-f})^T P_{f,s}^T y^f + \|P_{f,s}y^{-f}\|_2^2\right) \\
&= \|y\|_2^2 - \underbrace{\sum_{f=1}^{K}(y^f)^T P_{f,s}y^{-f}}_{\text{(I)}} - \underbrace{\sum_{f=1}^{K}(y^{-f})^T P_{f,s}^T y^f}_{\text{(II)}} + \underbrace{\sum_{f=1}^{K}(y^{-f})^T[(P_{f,s})^T P_{f,s}]y^{-f}}_{\text{(III)}}
\end{aligned} \tag{16}$$

Let $(\cdot)_f$ denote the block matrix with columns corresponding to the indices of f, so

$$(P_{f,s})^T P_{f,s} = \begin{bmatrix} (P_{f,s})_1^T(P_{f,s})_1 & \cdots & (P_{f,s})_1^T(P_{f,s})_K \\ \vdots & \ddots & \\ (P_{f,s})_K^T(P_{f,s})_1 & \cdots & (P_{f,s})_K^T(P_{f,s})_K \end{bmatrix} \tag{17}$$

$$P_{f,s}y^{-f} = \sum_{g\neq f}(P_{f,s})_g y^g, \quad (y^{-f})^T(P_{f,s})^T = \sum_{g\neq f}(y^g)^T(P_{f,s})_g^T. \tag{18}$$

We have found (16) is a quadratic form which we can write blockwise with the blocks given by folds. Dropping the $\|y\|_2^2$ term, we see that for block p the diagonal terms $(y^p)^T Q_{pp} y^p$ appear precisely in terms from (III) of the form $(y^p)^T(P_{h,s})_p^T(P_{h,s})_p y^p$ for $h \neq p$. So

$$Q^s_{ff} := \sum_{g\neq f}(P_{g,s})_f^T(P_{g,s})_f, \tag{19}$$

For $q \neq p$, the pq terms in (III) appear as $(y^p)^T(P_{h,s})_p^T(P_{h,s})_q y^q$ for all $h \notin \{p, q\}$. The only pq term in (I) occurs when $f = p$ and equals $(y^p)^T(P_{p,s})_q y^q$. Similarly, the only pq term in (II) occurs when $f = q$ and equals $(y^p)^T(P_{q,s})_p^T y^q$. Hence for $f \neq g$ we define

$$Q^s_{fg} := -(P_{f,s})_g - (P_{g,s})_f^T + \sum_{\substack{h=1 \\ h\notin\{f,g\}}}^{K} (P_{h,s})_f^T(P_{h,s})_g^T \tag{20}$$

Let y_K denote observations reordered according to the CV folds. We have shown that

$$\mathrm{RSS}(s) = \|y\|_2^2 + y_K^T Q^s y_K. \tag{21}$$

This allows us to characterize the cross-validation selection event.

Lemma 2.2. *Conditional on the models m_f fitted on each training set, the intersection of quadratic events*

$$\hat{s} = s \iff \bigcap_{\substack{r=1 \\ r \neq s}}^{S} \{y : y_K^T (Q^s - Q^r) y_K \leq 0\} \tag{22}$$

Since reordering the observations is accomplished by a permutation matrix $y_K = Py$, we conclude that the cross-validation selection procedure is characterized by quadratic inequalities.

Proposition 2.2. *Define E_j by including the events in (13) into $\mathbb{R}^n$, and let $E_r^{cv} := \{z : z^T P^T (Q^r - Q^s) P z \geq 0\}$, then*

$$M^f(y^f) = m_f \text{ for } f = 1, \ldots, K \text{ and } \hat{s} = s \iff y \in \bigcap_{j \in J} E_j \cap \bigcap_{r=1}^{S} E_r^{cv} \tag{23}$$

Propositions 2.1 and 2.2 allow us to adjust significance tests for variables included in the final model with sparsity level $\hat{s}$. Since the model selection events decompose as intersections of quadratic inequalities, given a distributional assumption on y we can condition to the selection event by simple operations of intersection and solving quadratics. We next discuss how the Lasso and other methods fit in this framework, and demonstrate some specific examples of inference conditional on this kind of model selection.

2.3. Extensions and applications

- Simplifying assumptions such as K dividing n and all folds having equal size were for notational ease only, and are not assumed in our software implementation.
- The model selection events of Lasso, forward stepwise, forward stepwise with groups, marginal screening, and many other methods fit in the quadratic framework (4). Any such method can be used to determine the events in (7) or (13).
- Similarly, the predictions of various methods, including the Lasso, forward stepwise, and others, result in various forms of the "hat matrices" (15). For the Lasso there are additional constant terms appearing in the fitted values which must be added into the RSS criterion. But these are constant on the model selection event, so the approach here still works and the only change is that the inequality defining E_λ^{cv} has additional constants.

- For the Lasso a grid of λ values can be used instead of the steps s of forward stepwise. Small modifications allow stagewise fitting such as `lar` (Efron et al., 2004) or `glmnet` (Friedman, Hastie and Tibshirani, 2010).
- The RSS criteria can be penalized in various ways to account for differences in model size if desired. For example, let $\lambda_{m,s} = 2|m_s|\log(n/K)$ be the BIC penalty with $|m_s|$ denoting the number of nonzero parameters in model m at step s and n/K the number of observations in the test set. Then we only need to add the constants $\lambda_{m,s} - \lambda_{m,r}$ to the left hand side of the corresponding quadratic inequalities.
- When σ^2 is unknown, there is a choice between using selective t or F tests or plugging in an estimate of σ. In the latter case, there are estimates that can be computed with cross-validation such as those discussed in Reid, Tibshirani and Friedman (2013).

3. Simulations

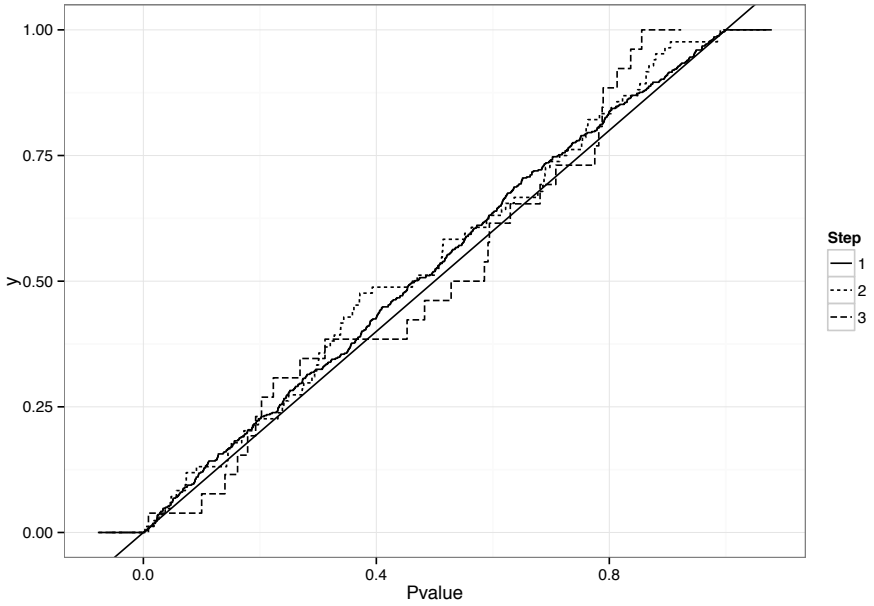

FIG 1. *Empirical CDF of selective p-values in a global null simulation with $n = 50$ observations of $p = 100$ independent Gaussian predictors. Cross-validation rarely chose models with sparsity larger than three, so only the first three steps are plotted.*

To demonstrate the applicability and power of this method we conducted simulations using forward stepwise as the model selection procedure as described in Section 2.2. Figures 1 and 2 show the empirical CDFs of selective p-values computed from truncated χ tests for the variables included in final models with

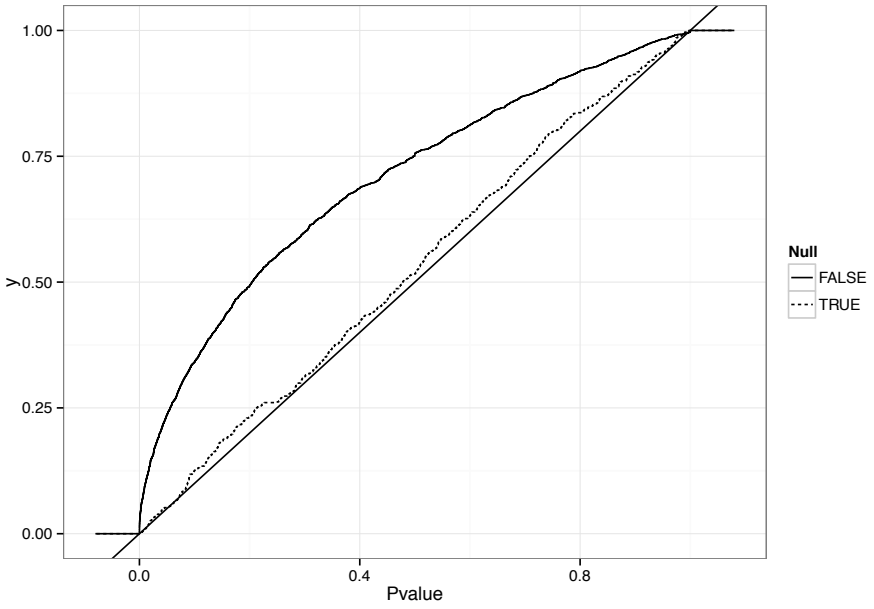

FIG 2. *Empirical CDF of selective p-values in a simulation similar to Fig. 1 but with true sparsity equal to five. The solid line shows that p-values corresponding to truly nonzero coefficients are small, so the test has power. The nonzero coefficients were all equal to* ± 1.

sparsity level determined by cross-validation. These figures show that significance tests for the variables which are null in the true model have the desired type 1 error control, and significance tests for the variables which are nonnull in the true model have reasonable power.

4. Discussion

The main drawback of our method is that it is computationally expensive. This cost is mostly due to the complicated geometry of the quadratic model selection regions (4). Important special cases, such as forward stepwise and the Lasso without any groups of variables, reduce to simpler polyhedral selection regions and this can be exploited by specialized implementations. This was not explored in the present work but will be included in a future version of the `selectiveInference` R package Tibshirani et al. (2015).

There are other limitations associated with the selective inference approach, but these are not particular to the present work on cross-validation. Perhaps the greatest of these limitations is that selective hypotheses may not be in correspondence with hypotheses about the true model when the model selection procedure performs poorly. By allowing the use of cross-validation—which is empirically known to perform quite well—in selective inference, the present work reduces the severity of this limitation.

Finally, the author is not aware of any previous work analyzing cross-validation through the quadratic form structure (19)–(21). Using this structure to obtain other new results on cross-validation unrelated to selective inference—for example, developing theory about the bias of minimum cross-validation error—is an area of ongoing work.

Acknowledgement: The author would like to thank Jonathan Taylor and Robert Tibshirani for helpful comments and suggestions.

References

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

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004). Least angle regression. *The Annals of statistics* **32** 407–499.

Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** 1–22.

Gross, S. M., Taylor, J. and Tibshirani, R. (2015). A Selective Approach to Internal Inference. *ArXiv e-prints*.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2015). Exact post-selection inference with the lasso. *Ann. Statist.* To appear.

Loftus, J. R. and Taylor, J. E. (2015). Selective inference in regression models with groups of variables. *ArXiv e-prints*.

Reid, S., Tibshirani, R. and Friedman, J. (2013). A study of error variance estimation in lasso regression. *arXiv preprint arXiv:1311.5274*.

Taylor, J. E., Loftus, J. R. and Tibshirani, R. J. (2015). Tests in adaptive regression via the Kac-Rice formula. *Ann. Statist.* To appear.

Tian, X., Loftus, J. R. and Taylor, J. E. (2015). Selective inference with unknown variance via the square-root LASSO. *arXiv preprint arXiv:1504.08031*.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2014). Exact Post-Selection Inference for Sequential Regression Procedures. *ArXiv e-prints*.

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J. and Reid, S. (2015). selectiveInference: Tools for Selective Inference R package version 1.1.1.