Post model-fitting exploration via a “Next-Door” analysis

Leying GUAN1* and Robert TIBSHIRANI2

1Department of Biostatistics, Yale University, New Haven, CT, U.S.A.
2Departments of Biomedical Data Sciences, and Statistics, Stanford University, Stanford, CA, U.S.A.

Key words and phrases: Feature selection; LOCO; model p-value; model selection; model score.

MSC 2010: Primary 62F03; secondary 62F07.

Abstract: We propose a simple method for evaluating the model that has been chosen by an adaptive regression procedure, our main focus being the lasso. This procedure deletes each chosen predictor and refits the lasso to get a set of models that are “close” to the chosen “base model,” and compares the error rates of the base model with that of nearby models. If the deletion of a predictor leads to significant deterioration in the model’s predictive power, the predictor is called indispensable; otherwise, the nearby model is called acceptable and can serve as a good alternative to the base model. This provides both an assessment of the predictive contribution of each variable and a set of alternative models that may be used in place of the chosen model. We call this procedure “Next-Door analysis” since it examines models “next” to the base model. It can be applied to supervised learning problems with $\ell_1$ penalization and stepwise procedures. We have implemented it in the R language as a library to accompany the well-known glmnet library.

Résumé: Les auteurs proposent une méthode simple pour l’évaluation des modèles choisis par une procédure de régression adaptative telle que le lasso, sur lequel ils se concentrent. Leur procédure consiste à retirer chaque prédicteur à tour de rôle et à réajuster le lasso afin d’obtenir un ensemble de modèles qui sont près du modèle de base. Ils comparent ensuite le taux d’erreur du modèle de base avec ceux de son voisinage. Lorsque le retrait d’une variable conduit à une baisse marquée de la puissance prédictive du modèle, le prédicteur est considéré comme indispensable. Sinon, il est jugé acceptable et peut servir de remplaçant pour le modèle de base. Cette approche permet à la fois de mesurer la contribution prédictive de chaque variable et de constituer un ensemble de modèles de remplacement. Les auteurs ont baptisé cette approche « l’analyse de la porte voisine » puisqu’elle consiste à examiner des modèles près du modèle de base. Celle-ci peut être appliquée aux problèmes d’apprentissage supervisé avec une pénalisation $\ell_1$, et des procédures pas-à-pas. Les auteurs ont implémenté leur méthode en R dans une bibliothèque de fonctions accompagnant la populaire bibliothèque glmnet.

1. INTRODUCTION

We consider the usual regression or classification situation: we have samples $(x_i, y_i), i = 1, 2, \ldots n$ where $x_i = (x_{i1}, \ldots x_{ip})^T$ and $y_i$ are the regressors and response for the $i$th observation. In regression, $y_i$ is quantitative while in classification it takes on one of $K$ discrete values. We will focus for now on the regression problem, but will discuss classification in Section 5.
We assume that an adaptive regression procedure has been fit to the data, and we want to assess the chosen model (base model). Our main focus in this paper is on the lasso, although procedures such as subset or stepwise regression may also be amenable to our approach. The lasso method solves the following problem (for simplicity, we have left out the intercept):

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n} \sum_i (y_i - x_i^T \beta)^2 + \lambda |\beta|,$$  \hspace{1cm} (1)

yielding a final model with sparse coefficients $\hat{\beta}$, for a sufficiently large value of $\lambda$. The data analyst is often interested in the importance of the selected predictors.

One way to measure the importance is to adopt a sub-model interpretation where we consider whether a predictor has a non-zero coefficient in the selected model. Conditional on the selected model, we can form post-selection $p$-values for the non-zero coefficients (Berk et al., 2013; Lee & Taylor, 2014; Fithian, Sun & Taylor, 2014; Lee et al., 2016; Tibshirani et al., 2016). Another way to measure its importance is to consider if the deletion of this predictor leads to significant deterioration in the predictive power given a training procedure. If the answer is “yes,” this predictor is indispensable. Otherwise, the new model trained without this feature is acceptable and may work as a substitute for the base model.

The two measures coincide when there is no feature selection. When $n \gg p$, we can fit a full regression model

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n} \sum_i (y_i - x_i^T \beta)^2.$$ 

If we restrict ourselves to the OLS regression only, the $p$-value that we obtain for each predictor reflects both (1) the significance of its coefficient being non-zero conditional in the current model, and (2) the deterioration in predictive power when the predictor is deleted and the model is refitted. When $p$ is large and especially when $p > n$, a full regression fit is not feasible, and the lasso is a popular approach for fitting. When we use the lasso penalty to select a model, however, these two criteria become different from each other. The recent progress in the field of post-selection inference has focused on the sub-model interpretation. In practice, researchers will sometimes be more interested in the second perspective.

Motivated by the discussion above, we propose to find and assess models “close” to the base model, a procedure that we call Next-Door analysis. The idea is as follows. We first fit the usual lasso, using cross-validation to choose $\lambda$. Then for each predictor in the support set, we remove that predictor and refit the lasso to all of the remaining predictors (not just the support set) using the chosen value of $\lambda$. This gives a nearby model (proximal model) corresponding to the deletion of each of the member in the support set. Finally, we examine and evaluate each of these nearby models. Algorithm 1 gives the details.

**Algorithm 1: Next-Door analysis for the lasso**

1. Fit the lasso with parameter $\lambda$ chosen by cross-validation. Let the solution be $\hat{\beta}(\lambda)$. Let $S$ be the active set where the coefficient in $\hat{\beta}(\lambda)$ is non-zero.
2. For each $j \in S$, solve the lasso problem with the coefficient for the $j$th predictor being fixed at 0:

$$\hat{\beta}(\lambda; j) = \arg\min_{\beta_j=0} \frac{1}{2n} \sum_i (y_i - x_i^T \beta)^2 + \lambda |\beta|.$$ \hspace{1cm} (2)

Let $d_j$ be the increase in the true validation error for this model relative to the base model.
3. Form an unbiased estimate of $d_j$ and test if predictor $j$ is indispensable: that is, test whether $d_j$ is positive.
Table 1: Prostate cancer results. The leftmost column shows the fitted model from the lasso, and the remaining columns show the nearby models corresponding to the removal of each predictor.

|          | base | lcavol | lweight | svi  | lcp  | lbph | pgg45 | age  |
|----------|------|--------|---------|------|------|------|-------|------|
| lcavol   | 0.64 | 0.69   | 0.70    | 0.59 | 0.65 | 0.63 | 0.62  |      |
| lweight  | 0.27 | 0.37   | 0.30    | 0.27 | 0.35 | 0.27 | 0.26  |      |
| svi      | 0.25 | 0.46   | 0.29    | 0.22 | 0.21 | 0.27 | 0.25  |      |
| lcp      | -0.12| 0.07   | -0.11   | -0.01| -0.14| -0.04| -0.11 |      |
| lbph     | 0.18 | 0.21   | 0.29    | 0.14 | 0.19 | 0.18 | 0.17  |      |
| pgg45    | 0.17 | 0.18   | 0.13    | 0.19 | 0.13 | 0.18 | 0.15  |      |
| age      | -0.08| -0.02  | -0.03   | -0.09| -0.07| -0.05| -0.07 |      |
| gleason  |      |        |         |      |      |      |       | 0.07 |
| cv_error | 0.61 | 0.90   | 0.65    | 0.64 | 0.62 | 0.61 | 0.63  | 0.60 |
| debiased_error | 0.62 | 0.94   | 0.66    | 0.66 | 0.63 | 0.62 | 0.62  | 0.62 |
| test_error | 0.51 | 0.87   | 0.49    | 0.56 | 0.50 | 0.50 | 0.47  | 0.53 |
| selection frequency | 1.00 | 1.00   | 0.96    | 0.78 | 1.00 | 0.88 | 0.74  |      |
| model p-value | 0.01 | 0.21   | 0.20    | 0.29 | 0.48 | 0.26 | 0.34  |      |
| model score | 0.01 | 0.21   | 0.21    | 0.37 | 0.48 | 0.30 | 0.45  |      |
| feature p-value | 0.00 | 0.01   | 0.02    | 0.23 | 0.05 | 0.07 | 0.28  |      |

As outlined in Algorithm 1, our test for indispensability considers the null hypothesis $H_0: d_j \leq 0$. It is challenging since the candidate models and the hypothesis are data adaptive and involve selections. One main task of this paper is to provide a good estimate of the $p$-value for the above test taking into consideration the selections.

Table 1 gives a preview of results from a Next-Door analysis. We apply it to a prostate cancer dataset (Friedman, Hastie & Tibshirani, 2001). The data consist of $n = 67$ training observations and 30 test observations. There are eight predictors. The response is the logarithm of the prostate specific antigen level for men who had prostate cancer surgery. Each column contains one set of model coefficients using a fixed training procedure. The columns corresponding to the proximal models are ordered according to their de-biased CV errors (from small to large). Details of the model $p$-value and model score for the “indispensability test” are provided in Section 2. The “selection frequency” is the proportion of times that the predictor is selected when the model fitting procedure is applied 50 times to bootstrap samples. The “feature $p$-value” is a post-selection $p$-value testing for non-zero coefficients. It is obtained using the R package `selectiveInference` (Lee et al., 2016). The feature $p$-values suggest that several predictors are significant, but only lcavol is indispensable considering the out-of-sample performance according to the model $p$-value and model score. For example, lweight is highly significant according to the feature $p$-value but not by the other two measures: the test error results suggest that the coefficients on other predictors can be adjusted to produce a model without much worse out-of-sample performance.

1.1. Related Work

Next-Door analysis measures the importance of a predictor by whether we can find a good model excluding this feature. It is closely related to the LOCO (leave-one-covariate-out) inference method.
(Rinaldo et al., 2016; Lei et al., 2018) and the variable importance measures used in random forest (Breiman, 2001). In the work of Rinaldo et al. 2016, a hold-out dataset is available. They do model selection, hypothesis selection and model fitting using only the training data. For each selected predictor, they coerce it to have a zero coefficient and rerun the model selection and training procedure. They then compare the performance of the original model and the new model in a hold-out validation set to evaluate its importance. They are able to do model-free inference conditional on the training data. Later, Markovic, Xia & Taylor (2017) suggest the use of marginalized LOCO. In the procedure of training a model with lasso penalty, instead of conditioning on the training data, they condition on a penalty being selected as well as the selected feature set $E$. They retrain the models with all data using OLS with features in $E$ and features in $E \setminus j$, and compare instead the prediction errors of these two models after marginalizing out the randomness in the training. Next-Door analysis essentially looks at a different type of marginalized LOCO parameter, without restricting ourselves to the selected feature set $E$. It is different from the work of Rinaldo et al. (2016) or Markovic, Xia & Taylor (2017) in the following ways:

1. Next-door analysis considers a different marginalization level. We marginalize out all randomness including the parameter tuning.
2. We do not have a hold-out dataset and we measure the importance of a feature by the test error of the CV models.
3. After the penalty is chosen with CV, we fix it when leaving out a predictor and retraining the model to loosely control the model complexity of Next-Door models to be similar to that of the original model. We can also vary this $\lambda$ as in Rinaldo et al. (2016). However, it does not seem to be necessary when we have marginalized out the randomness in the penalty picking step.

The answer of which marginalization level to consider should depend on how people make predictions in practice. For example, if we do not retrain the model with new data coming in, the LOCO conditional on the training data in Rinaldo et al. 2016 is more proper. However, if we repeat the whole training procedure including the parameter tuning, we may want a fully marginalized quantity.

Our proposal is also related to the low dimensional projection estimator (LDPE) (Zhang & Zhang, 2014; Zhu & Bradic, 2017; Yu, Bradic & Samworth, 2018). These estimators are concerned with the question of whether a predictor is important conditioning on all other predictors. To deal with the high dimensionality, LDPE is constructed using good initial model coefficient estimates and the part of a predictor that is “almost” orthogonal to other predictors. Our approach deals with high dimensionality through a different perspective and restricts ourselves to a small set of “accessible” models, which are models close to the base model in Next-Door analysis. Instead of looking at the coefficients, it looks directly at the prediction error. Another less related procedure is “stability selection” (Meinshausen & Bühlmann, 2010). This method identifies a set of “stable” variables that are selected with probability above a threshold by procedures like the lasso. Like the $p$-values from post-selection inference, even if a predictor is selected with reasonably high probability, it is still possible that we can find an alternative among the accessible models with similar prediction performance. For example, if we have two predictors that are identical and each of them is very important to the response without conditioning on the other, neither of them should be indispensable, but the selection probability will be around 0.5 for each of them.

The paper is organized as follows. In Section 2, we formalize how to test whether the difference in CV test errors $d_j$ is large with the full marginalization. We give details of the test method and give the definition of the model score in this section. In Section 3, we provide
intensive simulations to show the good performance of suggested methods. We apply Next-Door analysis to some real data examples in Section 4. In Section 5, we discuss the extension of Next-Door analysis to other settings.

2. TEST FOR INDISPENSABILITY WITH FULL MARGINALIZATION

In this section we give details of the “indispensability test” in Step 3 of Algorithm 1. Let \( \Lambda \) be the set of penalty parameters (in ascending order) that we consider at sample size \( n \) and suppose that we divide the data into finite folds \( \mathcal{V}_f \) with equal size. For any fixed penalty \( \lambda_k \), let \( S_k \) be the set of predictors selected. The CV errors for models trained with and without predictor \( j \) are \( Q_k \) and \( Q_k' \), defined as

\[
Q_k = \frac{1}{n} \sum_{i=1}^{n} Q_k(x_i, y_i), \quad Q_k' = \frac{1}{n} \sum_{i=1}^{n} Q_k'(x_i, y_i),
\]

where \( Q_k(x_i, y_i) \) and \( Q_k'(x_i, y_i) \) are the losses for the sample \((x_i, y_i)\) in CV for the base model and the next-door model without feature \( j \): let \( \hat{\beta}^v(\lambda_k) \) and \( \hat{\beta}^v(\lambda_k; j) \) be the fitted coefficients for fold \( v \) and at penalty \( \lambda_k \) from the base model and the next-door model respectively, and let \( \mathbb{1}_A \) be an indicator function for any event \( A \), then

\[
Q_k(x_i, y_i) = \sum_{v=1}^{V} (y_i - x_i^T \hat{\beta}^v(\lambda_k))^2 \mathbb{1}_{i \in \mathcal{V}_v}, \quad Q_k'(x_i, y_i) = \sum_{v=1}^{V} (y_i - x_i^T \hat{\beta}^v(\lambda_k; j))^2 \mathbb{1}_{i \in \mathcal{V}_v}.
\]

In practice, we will pick \( \lambda_k \) according to a criterion \( R \). In this section, we consider the case where we pick \( \lambda_k = \lambda_{k^*} \) to minimize the randomized validation error (to be discussed later). Other criterion could also be used. For example, one can use the CV one standard error rule (Friedman, Hastie & Tibshirani, 2001).

The chosen index \( k^* \) is a randomized quantity—if we do the selection with a different random seed, we can end up with a different penalty \( \lambda_{k^*} \). As we do not want to make judgement about predictor \( j \) based on a random quantity, we marginalize out the randomness in \( k^* \) and end up with the marginalized test error under the criterion \( R \). We let \( O_{k^*} \) be the event of selecting penalty index \( k^* \), and \((x, y)\) be an independent sample generated from their joint distribution. The test error after marginalization is defined as

\[
\text{Err}^R = \frac{1}{V} \sum_{k^*=1}^{m} \mathbb{E} \left[ \sum_{v=1}^{V} (y - x^T \hat{\beta}^v(\lambda_{k^*}))^2 \mathbb{1}_{O_{k^*}} \right],
\]

and

\[
\text{Err}^{(j)}^R = \frac{1}{V} \sum_{k^*=1}^{m} \mathbb{E} \left[ \sum_{v=1}^{V} (y - x^T \hat{\beta}^v(\lambda_{k^*}; j))^2 \mathbb{1}_{O_{k^*}} \right].
\]

We are interested in the hypothesis:

\[
H_0 : \text{Err}^{(j)}^R \leq \text{Err}^R \quad \text{vs.} \quad H_1 : \text{Err}^{(j)}^R > \text{Err}^R.
\]

The two events below prevent us from using the observed validation errors to do the test directly:

1. Selection event \( A_1 \) (model selection): The selected \( \lambda_{k^*} \) penalty achieves the smallest randomized CV errors among all \( \lambda_k \in \Lambda \).
2. Selection event \( A_2 \) (hypothesis selection): \( j \) is in the non-zero support \( S_{k^*} \).

DOI: 10.1002/cjs

The Canadian Journal of Statistics / La revue canadienne de statistique
To make the method more generalizable to complicated settings, we consider the events \( A_1 \) and \( A_2 \) separately. Intuitively, the event \( \{ j \in S_k \} \) should only have a small effect: the fact that the predictor \( j \) is selected will not typically have a big influence on the error of a refitted model that excludes this predictor, when the number of covariates is moderately large. However, the validation error obtained after selection event \( A_1 \) can be significantly biased (Tibshirani & Tibshirani, 2009).

We define the randomized cross-validation error and construct a de-biased test error estimate in Section 2.1. In Section 2.2, we describe the bootstrap \( p \)-value with the de-biased test error estimate considering only the event \( A_1 \). In Section 2.3, we propose a new importance measure called the model score, which uses the previous \( p \)-value to construct a quantity that controls the type I error after both selections \( A_1 \) and \( A_2 \). From a practical viewpoint, we recommend the use of the model score if the cost of falsely rejecting the null hypothesis is high; otherwise, the bootstrap \( p \)-value constructed in Section 2.2 usually works well and has higher power when signal detection is hard.

2.1. Randomized Cross-Validation Error and the De-Biased Error Estimate

For simplicity of notation, for a pre-fixed predictor \( j \), we let \( Q = (Q_1, \ldots, Q_m, Q'_1, \ldots, Q'_m) \) be the sequence of CV errors where the first \( m \) are from models trained with all predictors and the next \( m \) are from models trained with predictor \( j \) left out. Let \( \text{Err}_k = E[Q_k] \), and \( \text{Err} = (\text{Err}_1, \ldots, \text{Err}_m, \text{Err}'_1, \ldots, \text{Err}'_m) \). We define two sequences of randomized pseudo errors,

\[
\tilde{Q}^a(e, z) = Q + \frac{e}{\sqrt{n}} + \sqrt{\frac{\alpha}{n}} z, \quad \tilde{Q}^\frac{1}{a}(e, z) = Q + \frac{e}{\sqrt{n}} - \sqrt{\frac{1}{n\alpha}} z, \quad (3)
\]

where \( e \sim N(0, \gamma_1 \sigma_0^2 I) \), \( z \sim N(0, \Sigma + \gamma_1 \sigma_0^2 I) \) with \( \gamma_1 \) and \( \alpha \) being a positive constant and \( \sigma_0^2 \) being the smallest diagonal elements of \( \Sigma \). Here, \( \Sigma \) is an estimate of \( \Sigma \), the covariance of prediction errors using the population level coefficients: Let \( \beta(\lambda) = \arg \min \{ \frac{1}{2} \mathbb{E}(y - x^T \beta)^2 + \lambda | \beta | \} \), and \( \beta(\lambda; j) = \arg \min \beta_\lambda = \{ \frac{1}{2} \mathbb{E}(y - x^T \beta)^2 + \lambda | \beta | \} \). Then \( \Sigma \) is the covariance of the sequence of \( 2m \) prediction errors \((y - x^T \beta(\lambda_1)), \ldots, (y - x^T \beta(\lambda_m)), \ldots, (y - x^T \beta(\lambda_{m_1})), \ldots, (y - x^T \beta(\lambda_{m_2})), \ldots, (y - x^T \beta(\lambda_{m_2}))) \).

We choose the model index \( k^* \) to minimize the randomized validation errors \( \tilde{Q}^a_k(e, z) \) for \( k = 1, 2, \ldots, m \). In other words, we let the event \( O_{k^*} = (\tilde{Q}^a_{k^*}(e, z) \leq \tilde{Q}^a_k(e, z), \forall k = 1, \ldots, m \).

The first term \( \frac{e}{\sqrt{n}} \) is proposed by Rinaldo et al. 2016 to avoid the technical problem when applying the central limit theorem to the LOCO parameter in the sample splitting case. It is also used in Markovic, Xia & Taylor (2017) to make the randomized CV curves asymptotically normal with invertible covariance structure under suitable assumptions. The second terms, \( \sqrt{\frac{\alpha}{n}} z \) and \( \sqrt{\frac{1}{n\alpha}} z \), are introduced to make \( \tilde{Q}^a \) and \( \tilde{Q}^\frac{1}{a} \) marginally and asymptotically independent if \( \sqrt{n}(Q - \text{Err}) \) is asymptotically normal with covariance \( \Sigma \). This kind of parallel construction is proposed in Harris (2016). In their work, the authors estimate the prediction error for estimators like relaxed LASSO in the linear regression when the noise in the response \( y \) is homoscedastic Gaussian with variance \( \sigma^2 \). They construct two marginally independent responses \( y^a \) and \( y^\frac{1}{a} \) by adding noises \( \sqrt{\alpha} e \) and \( \frac{e}{\sqrt{\alpha}} \) to \( y \) with \( e \sim N(0, \sigma^2) \). Marginally, the prediction error estimate with \( y^\frac{1}{a} \) is unbiased for any selection performed using \( y^a \). A similar statement holds using the test or CV errors. Following this intuition, we estimate the test errors with Algorithm 2.1.
Suppose the subexponentiality assumption below holds and $Q$ satisfies Lemma 1. Then, as $n \to \infty$,

$$
\sqrt{n}(\mathbb{E}[\text{Err}]) - \sum_{k^* = 1}^m \text{Err}_{k^*} P(O_{k^*}) \to 0, \quad \sqrt{n}(\mathbb{E}[\text{Err}]) - \sum_{k^* = 1}^m \text{Err}_{k^* + m} P(O_{k^*}) \to 0.
$$

The behaviour of CV curves may not be asymptotically normal, however, under the consistency, subexponentiality and range assumptions below, the CV curves behave well enough such that the above test error estimates are almost unbiased. Throughout this section, we consider the setting where the dimension $p$ and the number of models $m$ are fixed, the features $x_j$ are standardized to have mean 0 and variance 1 for $j = 1, \ldots, p$, and $\mathbb{E}[xx^T]$ exists and is non-singular. Let $\beta^* = \arg\min_{\beta} \mathbb{E}(y - x^T \beta)^2$ and $\beta^*(j) = \arg\min_{\beta: \beta_j = 0} \mathbb{E}(y - x^T \beta)^2$ be the population level OLS solutions with and without feature $j$.

- **Consistency assumption:** For every $\lambda \in \Lambda_n$ and predictor index $j$ considered, we have that as $n \to \infty$,

$$
n\mathbb{E}[\|\hat{\beta}(\lambda) - \beta(\lambda)\|_2^4] \to 0, \quad n\mathbb{E}[\|\hat{\beta}(\lambda; j) - \beta(\lambda; j)\|_2^4] \to 0.
$$

- **Subexponentiality assumption:** $x_j$, $\forall j = 1, \ldots, p$, and $(y - x^T \beta^*)$ are sub-exponential.

- **Range assumption:** The smallest penalty considered satisfies $\lambda_1 < C n^{\frac{1}{4}}$ for a large enough constant $C$.

We have two remarks regarding the assumptions here: (1) The range assumption is mild and it only requires that we include at least one $\lambda$ that is of the order at most $O(n^{-\frac{1}{4}})$, while a suggested $\lambda$ to take is usually of order $O(n^{-\frac{1}{2}})$ when $p$ is fixed (Wainwright, 2009; Bickel, Ritov & Tsybakov, 2009). (2) Let $\tau$ be a constant. If $\sqrt{n} \lambda \to \tau$, $\sqrt{n}(\hat{\beta}(\lambda) - \beta^*)$ will converge to a well-defined distribution under mild assumptions, while if $\lambda \to 0$, $\hat{\beta}(\lambda)$ is consistent for estimating $\beta^*$ (Knight & Fu, 2000; Olive, 2017). The consistency assumption here is different and does not
assume a coverage to $\beta^*$. Instead, it is more about the variability of $\hat{\beta}(\lambda)$ at any given $\lambda$. For example, when $\lambda$ is very large and $\beta(\lambda) = 0$, then we require $\hat{\beta}(\lambda)$ to be around 0. Lemma 2 further reveals the difference between the consistency assumption and the results in Knight & Fu (2000) and Olive (2017).

Lemma 2. Under the subexponentiality assumption, for any $\lambda$, let $\beta = \arg\min \frac{1}{2} \| y - x^T \beta \|^2 + \lambda \| \beta \|$ and $\hat{\beta} = \arg\min \frac{1}{2m} \sum_{i=1}^m (y_i - x_i^T \beta)^2 + \lambda \| \beta \|$. If $\| \hat{\beta} \|_\infty \leq C$ for a constant $C$, then, $n \| \hat{\beta} - \beta \|_2 \to 0$ as $n \to \infty$.

Theorem 1 states that this procedure can successfully reduce the bias under the assumptions above. Define $K_1 := \{ k : \sqrt{n} \| \beta(\lambda_k) - \beta^* \|_2 < \infty \}$ and $K_2 := \{ m + k : \sqrt{n} \| \beta(\lambda_k; j) - \beta^*(j) \|^2 < \infty \}$. Theorem 1 says that when the signal size is not too large, the bias in our error estimates will be of order $o(\frac{1}{\sqrt{n}})$. When the signal size is very large (removing feature $j$ leads to a much worse model), the bias has a smaller order of that of the signal size $\Delta$.

Lemma 3. Suppose the consistency and subexponentiality hold. Let $\hat{\Sigma}$ be the sample covariance structure:

$$\hat{\Sigma}_{k,k'} = \frac{\sum_{i=1}^n (Q_k(x_i, y_i) - Q_k)(Q_{k'}(x_i, y_i) - Q_{k'})}{n}.$$ 

Then we have (1) $\| \hat{\Sigma} - \Sigma \|_\infty \to 0$ as $n \to \infty$, and (2) $\mathbb{E}[\sum_{k=1}^{2m} \hat{\Sigma}_{k,k}] \leq C$ for some constant $C$.

Remark 1. Lemma 3 suggests that we can use the sample covariance matrix estimate under assumptions in this paper. Another heuristic way to justify the use of sample covariance estimate is from a perspective conditional on the training model, which can be found in Lei (2017).

We give outlines of the proofs to Theorem 1, Lemmas 1, 2 and 3 in Appendix 1, and detailed proofs can be found in the Supplementary Material.

2.2. Bootstrap p-Value Approximation

Bootstrap (Bickel & Freedman, 1981; Efron & Tibshirani, 1994) is widely used to construct confidence intervals for quantities whose distributions may be hard to derive, for example, it has been used to construct confidence intervals for the model coefficients in the linear regression and generalized linear regression settings after variable selection (Tibshirani et al., 2018; Rathnayake & Olive, 2019; Watagoda & Olive, 2019). In the question we are considering, however, the bootstrap distributions of the selected mean test errors can be quite different from the original distributions, which can lead to poor performance (see Appendix 3). To alleviate this, we consider instead the debiased mean test errors and build a confidence interval for the quantity $T = (\hat{\text{Err}}^j - \text{Err}) - (\text{Err}^j - \text{Err})$. 

The Canadian Journal of Statistics / La revue canadienne de statistique DOI: 10.1002/cjs
Combining Theorem 1 and Lemma 3, we know that $T$ has expectation around 0 with a difference of order $o(\Delta + \frac{1}{\sqrt{n}})$ where $\Delta$ is the signal size defined as $\Delta = |\text{Err}^j R - \text{Err}^R|$. Let $T^*$ be the bootstrap version of $T$ by bootstrapping the errors. As a direct result of Lemma 1, the expectation of $T^*$ is around 0 with a difference of order $o(\frac{1}{\sqrt{n}})$. In other words, the first moments under the distributions of $T$ and $T^*$ are approximately equal, with a difference of order $o(\Delta + \frac{1}{\sqrt{n}})$.

Hence, we bootstrap $T$, and expect that the cumulative distribution function of $T^*$ to be close to that of $T$; let $F(x)$ be the CDF of $T$ and $F^*(x)$ be the CDF of $T^*$, we take the approximation that $F(x) \approx F^*(x)$. The $p$-value for the null hypothesis is then calculated as $p = 1 - F^*(\hat{\text{Err}}^j - \hat{\text{Err}})$. Some corrections can be introduced to improve the empirical performance of the type I error control. Here, we apply two modifications:

1. It is possible that the distributions of $\sqrt{n}T^*$ and $\sqrt{n}T$ are asymptotically degenerate. To account for this case, instead of looking at the empirical distribution of $T^*$, we look at the empirical distribution of $T^* + w$, where $w \sim \frac{\gamma}{\sqrt{n}} N(0, \sigma^2_0)$ for a small constant $\gamma$. The larger $\gamma$ is, the less power we will have and the more conservative we will be.

2. Let $\xi_k = \sqrt{n}(Q_k - \text{Err}_k)$ and, with a slight abuse of notation, let $\Sigma$ be its covariance, which is estimated by the sample covariance in practice. Let $\hat{\text{Err}}_1, \bar{Q}_1, \bar{\xi}_1$ be the means of the test error, validation error and $\xi_k$ for the first $m$ models respectively. Then we have

$$
\mathbb{E} \left[ \sum_{k=1}^m (Q_k - \bar{Q}_1)^2 \right] = \mathbb{E} \left[ \sum_{k=1}^m (\text{Err}_k - \text{Err}_1 + \frac{\xi_k}{\sqrt{n}} - \frac{\bar{\xi}_1}{\sqrt{n}})^2 \right] = \mathbb{E} \left[ \sum_{k=1}^m (\text{Err}_k - \text{Err}_1)^2 \right] + \frac{\sum_{k=1}^m \Sigma_{k,k} - \sum_{k,k'}^m \Sigma_{k,k'}}{nm}.
$$

We see that the bootstrap population has an inflated underlying test error dispersion due to noise. Let $Q^i$ be a vector of size $2m$. To match the average variability among the first $m$ models’ test errors, $k = 1, 2, \ldots, m$, we let

$$
Q^i_k = \sqrt{\frac{\sum_{k=1}^m (Q_k - \bar{Q}_1)^2 - (\sum_{k=1}^m \frac{\Sigma_{k,k}}{n} - \sum_{k,k'=1}^m \frac{\Sigma_{k,k'}}{nm})}{\sum_{k=1}^m (Q_k - \bar{Q}_1)^2}} (Q_k - \bar{Q}_1) + \bar{Q}_1.
$$

Similarly, let $\bar{Q}_2$ be the mean validation error for the $m$ models in the second half. To match the average variability among the second $m$ models’ test errors, for $k = m+1, m+2, \ldots, 2m$, we let

$$
Q^s_k = \sqrt{\frac{\sum_{k=m+1}^{2m} (Q_k - \bar{Q}_2)^2 - (\sum_{k=m+1}^{2m} \frac{\Sigma_{k,k}}{n} - \sum_{k,k'=m+1}^{2m} \frac{\Sigma_{k,k'}}{nm})}{\sum_{k=m+1}^{2m} (Q_k - \bar{Q}_2)^2}} (Q_k - \bar{Q}_2) + \bar{Q}_2.
$$

The mean-rescaled bootstrap is to do bootstrap in the population with the population mean $Q^s$ instead of $Q$. Let $Q(x_i, y_i)$ be $2m$ vector representing the loss for sample $i$. The mean-rescaled bootstrap generates samples from the mean-rescaled population:

$$
Q^s(x_i, y_i) \sim \{ Q(x_1, y_1), Q(x_2, y_2), \ldots, Q(x_n, y_n) \} - Q + Q^s. \quad (4)
$$
The $p$-value testing $H_0$ is constructed as $p = P(\hat{\text{Err}}^j - \text{Err} \leq T^* + w)$, where $T^*$ is the mean-rescaled bootstrap of the test statistics. We reject the null hypothesis when $p \leq \alpha$.

We provide simulations of this approximate $p$-value’s distribution in Appendix 2, the results show that an approximate $p$-value using bootstrap with the de-biased estimates is more uniform than that from the bootstrap using the observed CV errors.

**Bootstrap $p$-Value Approximation**

1. Input: Level $\alpha$, the errors $\{Q(x_i, y_i)\}$, $i = 1, 2, \ldots, n$, the number of bootstrap repetitions $B$, the extra noise level $\gamma_2$. By default, $B = 10,000$ and $\gamma_2 = 0.05$.
2. De-biased estimate: We apply Algorithm 2.1 to get the de-biased test error estimate for the $\hat{\text{Err}}$ and $\text{Err}^j$.
3. Let $\{Q(x_i, y_i)\}$, $i = 1, 2, \ldots, n$ be the rescaled bootstrap populations defined in the right-hand-side of (4).
4. Bootstrap: for each iteration $b$, we draw bootstrap samples from $\{Q(x_i, y_i)\}$, $i = 1, 2, \ldots, n$ and apply Algorithm 2.1 to the bootstrap samples. Let $\hat{\text{Err}}_b$ and $\text{Err}^j_b$ be the de-biased error estimates and $w_b \sim N(0, \frac{\gamma_2^2 \sigma^2}{n})$, then the $p$-value is calculated as

$$\hat{p} = \frac{\sum_{b=1}^B 1(\hat{\text{Err}}^j_b - \text{Err}_b - (Q^i_{\text{err}} - Q^i_{\text{err}}) + w_b \geq \hat{\text{Err}}^j - \text{Err})}{B}. $$

where $Q^i_{\text{err}}$ and $Q^i_{\text{err}} - Q^i_{\text{err}}$ are the test errors under criterion $R$ using the mean-rescaled bootstrap population:

$$Q^i_{\text{err}} = E\left[\sum_{k^* = 1}^m Q_{k^*} 1_{O_{k^*}}\right], \quad Q^i_{\text{err}} - Q^i_{\text{err}} = E\left[\sum_{k^* = 1}^m Q_{k^* + m} 1_{O_{k^*}}\right].$$

We reject the null hypothesis for the predictor $j$ if $\hat{p} \leq \alpha$.

2.3. Model Score: A Conservative Measure of Importance

In this section, we consider an additional post-processing step to deal with the selection event $A_2$ and to guard against being overly optimistic. This post-processing step accounts for the over-optimism due to selection event $A_2$ by discounting the importance of a predictor based on how frequently it is selected by the model. For a feature $j$, let $p_j$ be its $p$-value considering only the selection event $A_1$. Let $\{j \in S_{k^*}\}$ represent the event that $j$ is in the selected model with penalty $\lambda_{k^*}$. Then, the model score for predictor $j$ is defined as $s_j = \frac{p_j}{\gamma_j}$, where

$$\gamma_j := \sum_{k^* = 1}^m p(j \in S_{k^*}, O_{k^*})$$

is the average selection frequency for predictor $j$. Recall that $O_{k^*}$ is the event that we select the model with penalty $\lambda_{k^*}$. For example, if we select the penalty minimizing the randomized CV error, then $O_{k^*} = \{\hat{Q}_{k^*}^n \leq Q_{k^*}^n \forall k = 1, \ldots, m\}$.

- Selection frequency assumption: as $n \to \infty$, the selection frequency of a predictor $j$ converges to a constant $\lim_{n \to \infty} \gamma_j = c_j \in [0, 1]$.

We only consider predictors whose selection frequency has a limit greater than 0. For those predictors with a non-vanishing selection frequency, we can control the type I error at level $\alpha$ asymptotically by rejecting only $s_j < \alpha$. 

The Canadian Journal of Statistics / La revue canadienne de statistique DOI: 10.1002/cjs
Theorem 2. Suppose that the selection frequency assumption holds and the predictor $j$ is a predictor with $c_j > 0$. Let $p_j$ be a $p$-value such that $\lim_{n \to \infty} P_{H_0}(p_j \leq \alpha) \leq \alpha$ for any fixed $\alpha$. Let $s_j = \frac{p_j}{r_j}$, then we have $\lim_{n \to \infty} P_{H_0}(s_j \leq \alpha | j \in S_{k^*}) \leq \alpha$.

Proof. By definition, $P(j \in S_{k^*}) = \gamma_j$, hence,

$$P_{H_0}(s_j \leq \alpha | j \in S_{k^*}) = \frac{P_{H_0}(p_j \leq \gamma_j \alpha, j \in S_{k^*})}{\gamma_j} \leq \frac{P_{H_0}(p_j \leq \gamma_j \alpha)}{\gamma_j}.$$

We take the limit of the above inequality and apply the Slutsky’s theorem to conclude the proof:

$$\lim_{n \to \infty} \frac{P_{H_0}(p_j \leq \gamma_j \alpha)}{\gamma_j} = \frac{P_{H_0}(p_j \leq \gamma_j \alpha)}{\gamma_j} \leq \alpha.$$

The denominator of the model score $s_j$ is $\gamma_j$, the frequency of a predictor being selected using criterion $R$, which is also used by stability selection (Meinshausen & Bühlmann, 2010). In practice, we can estimate $\gamma_j$ by doing a paired bootstrap of $(X, Y)$, refitting the models, and picking the $k^*$ using the new errors. We then estimate the frequency of predictor $j$ being selected in those models. Also, we will set a small cut-off, say, 0.05, on the observed selection frequency $\gamma_j$ and we do not reject a predictor if $\gamma_j$ is smaller than that.

3. SIMULATIONS

In this section, we evaluate the performance of our proposal in the linear regression setting and compare them to predictor $p$-values from post-selection inference and a naive approach neglecting all selections when looking at the model errors. We consider both model $p$-value, the de-biased approach neglecting the selection event $A_2$, and model score, the de-biased approach using model score to account for $A_2$. At any given level $\alpha$, a rejection using the model score is the same as a rejection using a $p$-value: we reject a hypothesis if its score is smaller than a given level $\alpha$. The $p$-values using the naive approach is referred to as model $p$-value(naive).

For the post-selection inference approach, we consider the post-selection feature $p$-value feature $p$-value from the selectiveInference package (Lee et al., 2016). We also consider a new post-selection approach that measures the model’s prediction power neglecting the selection event $A_2$, referred to as model $p$-value (post-selection) and the details are given in Appendix 3. We include the later to support our claim that the event $A_2$ does not have a significant influence.

We generate $n = 100$ observations from a linear model

$$y_i = \beta_0 + \sum_{j=1}^{p} X_{ij} \beta_j + Z_i,$$

for different dimensions $p$ and sparsity levels $s$. When $s \neq 0$, we set $\beta = \left( \frac{2}{1}, \frac{2}{2}, \ldots, \frac{2}{s}, 0, \ldots, 0 \right)$. In our simulations, we consider $Z_i \sim N(0, 1)$ and $Z_i \sim t_5$, and for a given $(s, p)$, we examine the following four simulation settings:

Orthogonal design: Let $X$ be standard Gaussian predictors.

Redundant design I: This is a setting designed specifically for Next-Door analysis. The design matrix is in a way such that almost no predictor is indispensable. Let the first half predictors $X_{1 \frac{p}{2}}$ and $W$ be standard Gaussian predictors with length $\frac{p}{2}$, and the second half predictors $X_{\left(\frac{p}{2}+1\right):p} = 0.95X_{\left(\frac{p}{2}+1\right):\frac{p}{2}} + 0.05W$.

Correlated design: Let $X$ be Gaussian predictors with variance 1 and corr$(X_j, X_k) = 0.5$.

DOI: 10.1002/cjs

The Canadian Journal of Statistics / La revue canadienne de statistique
Redundant design II: Let the first half predictors $X_{1:n2}$ be Gaussian with variance 1 and $\text{corr}(X_j, X_k) = 0.5$, and $W$ be standard Gaussian predictors with length $n2$. The second half predictors $X_{(n2+1):n2}$ are given by $0.95X_{1:n2} + 0.05W$.

We repeat the above experiments for 100 times. Empirical type I errors with a targeted coverage being 90% are given in Table 2. For the type I error calculation, we consider predictors with index $j \geq s + 1$ in the non-redundant case and $j = \frac{n2}{2} + 1, \ldots, \frac{n2}{2} + s$ in the redundant case. The power curves under a non-redundant design with $(p, s) = (10, 5), (10, 10), (400, 5), (400, 30)$ are given in Figure 1.

In our simulations, both the bootstrap model p-value and post-selection model p-value considering only $A_1$ have reasonable performance in controlling the type I error on average. The model scores are conservative and perform well in controlling the type I error. The naive approach and feature p-value cannot control the type I error, as expected. Figure 1 shows that

1. Comparing the bootstrap model p-value for the marginalized test error and the post-selection model p-value conditional on the penalty selected, there is a loss in power for the latter.
2. Comparing model p-value with the feature p-value, conditioning on the whole selected feature set $E$ can lead to a dramatic power loss in high dimensional settings.

In practice, the model p-value that neglects the selection $A_1$ is generally well-behaved. It also has less computational cost and higher power in extremely small signals. However, the model score approach may be preferred in cases where exact type I error control is essential. Code for the simulations can be found in the Supplementary Material.

4. REAL DATA APPLICATIONS

In this section, we provide two more real data examples. The second example uses the HIV data from Rhee et al. 2003 where the authors studied six nucleoside reverse transcriptase inhibitors that are used to treat HIV-1. We take the measurement of one of the inhibitors 3TC as the response, and 240 mutation sites as the predictors. There are 1,073 samples in this experiment. We randomly split the samples into 800 training samples and 273 test samples. The mutation site p184 is special, and it has prediction power dominating all other sites. Results are given in Table 3. The original randomized model selected 21 predictors. For the sake of space, we do not include in the table those predictors whose model p-value and feature p-value are both greater than 0.2, and whose test errors for their proximal models are no greater than that for the original model, which leaves 15 predictors for use. With a p-value cut-off being 0.1, four predictors, p184, p65, p215 and p69, are found indispensable. For the rows, we do not show predictors that only appear in the models deleting p184, p65, p215 or p69.

As a third example, we apply Next-Door analysis to a gastric cancer dataset, consisting of measurements on $p = 2,200$ proteins, from each of 12,480 pixels (observations) obtained from 14 patients (Eberlin et al., 2014). In this example, instead of selecting the model with the smallest randomized CV error, we use the CV one standard error rule. The CV folds are the same as the patients’ ids. The outcome is cancer ($Y = 1$) versus normal ($Y = 0$), and we fit a lasso-regularized logistic regression. The errors are based on the deviance from the fitted model. The results are shown in Table 4. We select 19 proteins in the base model and 28 proteins in total are selected for all 19 proximal models. Among the 19 proteins in the base model, we keep only the 15 which have at least one p-value no greater than 0.05. Among the 28 proteins, we keep in the rows only the 20 proteins whose coefficients’ magnitude is at least 0.05, to save the space. The model p-values suggest that the first 6 proteins (#487, #476, #607, #431, #1049, #552) can be important to the models’ predictive power with a p-value cut-off being 0.1. The protein #1509 is on the boundary (model p-value being 0.127), it might also be important as its
Table 2: Empirical type I error with pre-specified level $\alpha = 0.1$ under four different experiment settings and different $(p, s)$ combinations. The columns with names $M_{1i}$ for $i = 1, 2, 3$ show results from the simulations where $p = 10$ and $s = 0, 5, 10$, respectively. Similarly, the columns with name $M_{2i}$ for $i = 1, 2, 3$ show results from the simulations where $p = 400$ and $s = 0, 5, 30$, respectively. The column names “O,” “RI,” “C” and “RII” represent “Orthogonal Design,” “Redundant Design I,” “Correlated Design” and “Redundant Design II” respectively. Also, we let “$m_p$-value,” “$m$-score” and “$f_p$-value” refer to “model $p$-value,” “model score” and “feature $p$-value” respectively, and use “post” to represent “post selection”.

| $Z_i \sim N(0, 1)$ | $M_{11}$ | $M_{12}$ | $M_{13}$ | $M_{21}$ | $M_{22}$ | $M_{23}$ |
|---------------------|----------|----------|----------|----------|----------|----------|
|                     | O  | RI | C   | RII  | O  | RI | C   | RII  | O  | RI | C   | RII  |
| $m_p$-value(naive)  | 0.04 | 0.08 | 0.09 | 0.11 | 0.01 | 0.13 | 0.04 | 0.14 | 0.21 | 0.26 | 0.22 | 0.20 |
| $m_p$-value         | 0.05 | 0.07 | 0.06 | 0.06 | 0.03 | 0.10 | 0.04 | 0.07 | 0.13 | 0.11 | 0.13 | 0.10 |
| $m$-score           | 0.03 | 0.03 | 0.04 | 0.05 | 0.02 | 0.06 | 0.04 | 0.06 | 0.08 | 0.10 | 0.09 | 0.10 |
| $m_p$-value(post)   | 0.08 | 0.12 | 0.09 | 0.09 | 0.08 | 0.05 | 0.10 | 0.13 | 0.12 | 0.10 | 0.13 | 0.10 |
| $f_p$-value         | 0.12 | 0.15 | 0.16 | 0.19 | 0.09 | 0.56 | 0.13 | 0.56 | 0.12 | 0.13 | 0.11 | 0.18 |

| $Z_i \sim t_5$ | $M_{11}$ | $M_{12}$ | $M_{13}$ | $M_{21}$ | $M_{22}$ | $M_{23}$ |
|-----------------|----------|----------|----------|----------|----------|----------|
|                 | O  | RI | C   | RII  | O  | RI | C   | RII  | O  | RI | C   | RII  |
| $m_p$-value(naive)  | 0.05 | 0.07 | 0.04 | 0.06 | 0.01 | 0.12 | 0.01 | 0.10 | 0.25 | 0.25 | 0.24 | 0.21 |
| $m_p$-value         | 0.04 | 0.07 | 0.08 | 0.07 | 0.04 | 0.09 | 0.05 | 0.10 | 0.11 | 0.08 | 0.11 | 0.10 |
| $m$-score           | 0.02 | 0.04 | 0.06 | 0.05 | 0.04 | 0.08 | 0.04 | 0.08 | 0.09 | 0.09 | 0.10 | 0.09 |
| $m_p$-value(post)   | 0.09 | 0.09 | 0.10 | 0.09 | 0.10 | 0.12 | 0.11 | 0.14 | 0.11 | 0.15 | 0.13 | 0.12 |
| feature $p$-value   | 0.11 | 0.15 | 0.10 | 0.12 | 0.09 | 0.55 | 0.12 | 0.60 | 0.12 | 0.13 | 0.11 | 0.14 |

DOI: 10.1002/cjs
FIGURE 1: Power curves for six approaches with \((p,s) = (10,5), (10,10), (400,5), (400,30)\). The solid blue and dashed red curves are the power curve using the bootstrap \(p\)-value and its model score respectively. The solid green curve is for the model that uses the post-selection model \(p\)-value without considering \(A_2\). The solid black curve uses the \(t\)-test and nominal CV error that ignores all selection events. The dashed black curve uses post-selection for features.

proximal model has de-biased CV error larger than that of three other selected proteins. In this example, because of the heterogeneity of \((x, y)\) from different patients (14 patients in the training data and 5 patients in the test data), the alignment between the test error and the CV error is not as good as the previous two examples. Data and code used in this section can be found in the Supplementary Material.
Table 3: HIV dataset results. The leftmost column shows the fitted model lasso, and the remaining columns show the proximal model corresponding to the removal of each predictor. For a better visualization, we have scaled up both the coefficients, the test error and test error estimations by a factor of 10, and keep only 2 digits after the decimal. The rowname “p54-” means that the values in this row are the negative of the coefficients ($\times 10$) of p54.

|       | Base | p184 | p65 | p215 | p69 | p228 | p33 | p172 | p75 | p54 | p210 | p67 | p15 | p90 | p151 | p62 |
|-------|------|------|-----|------|-----|------|-----|------|-----|-----|------|-----|-----|-----|------|-----|
| p184  | 9.34 | 9.33 | 9.36 | 9.34 | 9.34 | 9.34 | 9.34 | 9.33 | 9.34 | 9.34 | 9.34 | 9.34 | 9.34 | 9.34 | 9.34 |
| p65   | 1.10 | 0.83 | 1.05 | 1.10 | 1.10 | 1.10 | 1.10 | 1.10 | 1.10 | 1.11 | 1.10 | 1.10 | 1.10 | 1.11 | 1.12 |
| p215  | 0.82 | 1.62 | 0.63 | 0.83 | 0.82 | 0.82 | 0.81 | 0.84 | 0.93 | 0.82 | 0.82 | 0.81 | 0.83 |
| p69   | 0.23 | 0.23 | 0.21 | 0.25 | 0.26 | 0.23 | 0.24 | 0.24 | 0.23 | 0.28 | 0.23 | 0.23 | 0.24 |
| p228  | 0.11 | 0.13 | 0.06 | 0.15 | 0.17 | 0.11 | 0.12 | 0.11 | 0.12 | 0.05 | 0.11 | 0.12 | 0.11 |
| p33   | 0.04 | 0.10 | 0.03 | 0.05 | 0.03 | 0.04 | 0.05 | 0.04 | 0.04 | 0.03 | 0.04 | 0.04 | 0.04 |
| p172  | 0.02 | -0.34 | 0.01 | 0.04 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | 0.02 | 0.03 | 0.02 | 0.02 |
| p75   | 0.09 | 0.04 | 0.04 | 0.10 | 0.10 | 0.09 | 0.09 | 0.08 | 0.10 | 0.13 | 0.09 | 0.10 | 0.09 |
| p54-  | 0.11 | 0.10 | 0.10 | 0.12 | 0.11 | 0.11 | 0.11 | 0.13 | 0.09 | 0.11 | 0.12 | 0.11 | 0.12 |
| p210  | 0.15 | 0.19 | 0.19 | 0.24 | 0.14 | 0.17 | 0.15 | 0.16 | 0.16 | 0.18 | 0.15 | 0.15 | 0.15 |
| p67   | 0.38 | 0.41 | 0.39 | 0.56 | 0.43 | 0.35 | 0.38 | 0.39 | 0.39 | 0.38 | 0.38 | 0.38 | 0.37 |
| p115  | 0.00 | 0.85 | 0.09 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
| p90   | 0.09 | 0.44 | 0.09 | 0.10 | 0.08 | 0.10 | 0.09 | 0.10 | 0.10 | 0.09 | 0.10 | 0.09 | 0.09 |
| p151  | 0.10 | 0.13 | 0.08 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 | 0.11 | 0.10 | 0.10 | 0.10 | 0.10 |
| p62   | 0.10 | 0.51 | 0.25 | 0.14 | 0.11 | 0.10 | 0.10 | 0.10 | 0.12 | 0.09 | 0.07 | 0.10 | 0.10 |
| p25   | 0.01 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| p125  | 0.06 | -0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| cv    | 0.62 | 8.28 | 0.78 | 0.64 | 0.63 | 0.62 | 0.62 | 0.62 | 0.62 | 0.62 | 0.62 | 0.62 | 0.62 |
| debias | 0.63 | 8.47 | 0.78 | 0.65 | 0.64 | 0.64 | 0.63 | 0.63 | 0.63 | 0.63 | 0.63 | 0.63 | 0.62 |
| test  | 0.63 | 8.72 | 0.85 | 0.65 | 0.64 | 0.63 | 0.63 | 0.63 | 0.63 | 0.63 | 0.63 | 0.63 | 0.63 |
| freq  | 1.00 | 1.00 | 1.00 | 1.00 | 0.90 | 0.66 | 0.84 | 0.74 | 0.68 | 1.00 | 1.00 | 0.80 | 1.00 |
| m_p-value | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.67 | 0.15 | 0.50 | 0.51 | 0.23 | 0.69 | 0.14 | 0.84 |
| m_score | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.77 | 0.23 | 0.60 | 0.68 | 0.34 | 0.69 | 0.14 | 1.05 |
| f_p-value | 0.00 | 0.000 | 0.01 | 0.02 | 0.13 | 0.28 | 0.44 | 0.22 | 0.01 | 0.12 | 0.03 | 0.91 | 0.09 | 0.10 | 0.11 |
Table 4: Gastric cancer data. The leftmost column shows the fitted model lasso, and the remaining columns show the proximal model corresponding to the removal of each predictor.

|   | Base | 487 | 476 | 607 | 1509 | 431 | 1049 | 552 | 1648 | 608 | 1374 | 606 | 1453 | 423 | 894 | 171 |
|---|------|-----|-----|-----|------|-----|------|-----|------|-----|------|-----|------|-----|-----|-----|
| 487| 0.578| 0.555| 0.584| 0.589| 0.639| 0.618| 0.563| 0.576| 0.583| 0.580| 0.591| 0.582| 0.569| 0.560| 0.607|
| 476| 0.339| 0.272| 0.356| 0.347| 0.354| 0.338| 0.397| 0.333| 0.352| 0.334| 0.326| 0.335| 0.347| 0.364| 0.347|
| 607| 0.165| 0.188| 0.185| 0.154| 0.162| 0.184| 0.186| 0.165| 0.192| 0.166| 0.194| 0.167| 0.175| 0.182| 0.168|
| 1509-| 0.206| 0.219| 0.196| 0.223| 0.207| 0.201| 0.207| 0.204| 0.206| 0.205| 0.232| 0.212| 0.193| 0.205|
| 431| 0.244| 0.527| 0.258| 0.242| 0.278| 0.278| 0.220| 0.245| 0.238| 0.241| 0.229| 0.265| 0.249| 0.238| 0.360| 0.220|
| 1049| 0.198| 0.314| 0.193| 0.213| 0.200| 0.181| 0.237| 0.205| 0.208| 0.199| 0.207| 0.196| 0.200| 0.259| 0.201|
| 552| 0.200| 0.174| 0.241| 0.213| 0.196| 0.201| 0.227| 0.201| 0.215| 0.205| 0.200| 0.202| 0.205| 0.194| 0.204|
| 1648| 0.064| 0.041| 0.047| 0.066| 0.067| 0.055| 0.081| 0.068| 0.061| 0.072| 0.076| 0.064| 0.064| 0.007| 0.058|
| 1038| 0.144| 0.091| 0.137| 0.127| 0.168| 0.201| 0.188| 0.129| 0.159| 0.137| 0.154| 0.162| 0.155| 0.147| 0.155| 0.146|
| 551| 0.061| 0.131| 0.033| 0.081| 0.073| 0.087| 0.054| 0.114| 0.065| 0.069| 0.063| 0.066| 0.061| 0.062| 0.073| 0.058|
| 608| 0.083| 0.101| 0.100| 0.115| 0.080| 0.080| 0.098| 0.112| 0.082| 0.087| 0.089| 0.084| 0.086| 0.081| 0.084|
| 475| 0.085| 0.078| 0.294| 0.071| 0.051| 0.063| 0.051| 0.047| 0.098| 0.077| 0.096| 0.105| 0.083| 0.145| 0.095| 0.075|
| 1596| 0.021| 0.074| 0.009| 0.011| 0.108| 0.001| 0.022| 0.057| 0.015| 0.033| 0.014| 0.013| 0.019| 0.082| 0.033|
| 1374| 0.050| 0.064| 0.035| 0.052| 0.048| 0.031| 0.053| 0.063| 0.056| 0.057| 0.049| 0.048| 0.038| 0.049|
| 606| 0.098| 0.173| 0.065| 0.160| 0.093| 0.128| 0.122| 0.100| 0.109| 0.108| 0.107| 0.095| 0.096| 0.039| 0.092|
| 1453-| 0.043| 0.084| 0.028| 0.050| 0.171| 0.055| 0.036| 0.053| 0.043| 0.046| 0.043| 0.039| 0.050| 0.055| 0.043|
| 423| 0.088| 0.019| 0.113| 0.113| 0.112| 0.078| 0.090| 0.110| 0.088| 0.093| 0.085| 0.086| 0.092| 0.061| 0.092|
| 894| 0.242| 0.166| 0.267| 0.257| 0.226| 0.298| 0.294| 0.240| 0.228| 0.243| 0.240| 0.226| 0.245| 0.234| 0.244|
| 171| 0.035| 0.209| 0.051| 0.040| 0.043| 0.031| 0.002| 0.043| 0.046| 0.030| 0.036| 0.034| 0.029| 0.034| 0.039| 0.040|
| 898| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059| 0.059|

The Canadian Journal of Statistics / La revue canadienne de statistique
DOI: 10.1002/cjs
5. EXTENSIONS

5.1. Generalization to Other Supervised Learning Algorithms

The methods proposed here can be extended in a straightforward manner to Cox’s proportional hazards model and the class of generalized linear models where the outcome $Y$ depends on a parameter vector $\eta$:

$$\eta = X\beta.$$  \hfill (5)

In this case, we have the penalized negative log likelihood as the objective function

$$J(\beta) = -\ell(\beta) + \lambda|\beta|.$$  \hfill (6)

The event $A_1$ can be characterized using the corresponding new CV loss we are interested in, and the selection frequency $\gamma_j$ remains unchanged. As a result, the $p$-value neglecting the selection $A_2$ and the model score are both easily obtained in more complicated scenarios where we do not know how to characterize the selection for event $A_2$ in an efficient way, even for a black box model. The Gastric cancer dataset is an example where we apply the Next-door analysis to a classification problem.

We can also apply the idea of next-door analysis to non-linear models and models with interactions, for example, lasso with interaction terms and random forests (RF), by retraining next-door models without a feature $j$. In this case, however, the practitioners may find the following two things undesirable: (1) The resulting next-door models are more likely to be far apart from the base model due to the complicated structure, and as a result, it may be difficult to make comparisons. (2) Methods like RF do not perform feature selection and it is computationally expensive to rerun the whole training procedure from scratch removing every single feature. One potential future direction for the next-door analysis is to apply a structural constraint to next-door models without a feature $j$ and reduce the model discrepancy between the base model and the next-door models, as well as the computational cost. For example, in RF, one constraint we can investigate is to fix the locations of selected variables in the base model which are not feature $j$. As we go down each tree in the base model, when we encounter feature $j$, we will replace it with another feature, otherwise, we use the current feature in the base model. For RF, such a constrained version of next-door analysis will be a balance between the variable importance, which fixes the prediction model and permutes a feature $j$, and an unconstrained next-door analysis.

In this paper, we considered some assumptions under which the asymptotic normality holds for the randomized CV curve. We observe that such a normality assumption usually holds approximately in practice, and the bootstrap $p$-value itself is also usually robust.

5.2. A Model with Better Out-Of-Sample Performance

The model $p$-value and model score can serve as an alternative feature importance measure even when we consider features only in the current selected feature set $E$. It provides a different ordering of feature importance compared with $p$-value, correlation or partial correlation with the response. It can work better sometimes in practice as it considers the out-of-sample error directly. We provide examples using the prostate data, HIV data and the gastric cancer data. We consider the selected feature set in the first stage and retrain the model using the training data with OLS/logistic regression. We build a sequence of nested models in which we add feature one by one according to their model $p$-value, model score and feature $p$-value. For the prostate dataset and gastric cancer dataset, we start from models containing one feature. For the HIV dataset, we start from models containing two features as the test errors are much larger for models with only one feature compared with the others. In Figure 2, we evaluate the models...
FIGURE 2: Test errors for nested model sequences created based on model $p$-value, model score and feature $p$-value. The vertical dashed line is where we want to stop based on the model $p$-values as described in Section 4.

out-of-sample performance in the test set as a function of the number of features added. The vertical dashed line is the number where we want to stop based on the model $p$-value. In all three cases, the model $p$-values produced more sparse models with near optimal performance (smallest test errors achieved using the nested procedure).

6. DISCUSSION

Our post-fitting procedure Next-Door analysis gives insights into predictor indispensability and offers nearby alternative models. Our proposal shifts the focus from coefficients to models: having selected a model from the data, we look for alternative models that omit each predictor, and yet have validation error similar to the base model. The model performance is considered marginalizing out the parameter turning and randomness in the training. We present a bootstrap approach based on the de-biased test error estimate for a pre-fixed hypothesis. We also propose a simple concept called model score which takes into account the hypothesis selection by considering its selection frequency. By considering the hypothesis selection and model selection separately in this paper, we can easily deal with more complicated model selection and hypothesis selection events. An R package called “nextdoor” for performing Next-door analysis will soon be available on the public CRAN repository.

Next-Door analysis can also be used in cases where one wants to examine the removal of a set of predictors. In the case where the users have in mind which $k$ predictors they do not want to use after looking at the fitted model, they can simply remove those predictors and all analyses will still carry through. In general, however, it is not practical to enumerate all different combinations of $k$ predictors.

ACKNOWLEDGEMENTS

We would like to thank Zhou Fan, Ryan Tibshirani and Jonathan Taylor for the helpful discussions and feedback. We would also like to thank the referees and editors for carefully reading the paper and providing useful suggestions on the presentations. Robert Tibshirani was supported by NIH grant 5R01 EB001988-16 and NSF grant 9 DMS1208164.

BIBLIOGRAPHY

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. The Annals of Statistics, 41, 802–837.

Bickel, P. J. & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. The Annals of Statistics, 9, 1196–1217.
Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37, 1705–1732.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

Eberlin, L. S., Tibshirani, R. J., Zhang, J., Longacre, T. A., Berry, G. J., Bingham, D. B., Norton, J. A., Zare, R. N., & Poultsides, G. A. (2014). Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 2436–2441.

Efron, B. & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press.

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

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning. *Springer Series in Statistics*, Springer, New York.

Harris, X. T. (2016). Prediction error after model search. arXiv preprint arXiv:1610.06107.

Knight, K. & Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics*, 28, 1356–1378.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44, 907–927.

Lee, J. D. & Taylor, J. E. (2014). Exact post model selection inference for marginal screening. *Advances in Neural Information Processing Systems*; p. 136–144.

Lei, J. (2017). Cross-validation with confidence. arXiv preprint arXiv:1703.07904.

Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113, 1094–1111.

Markovic, J., Xia, L. u.c.y., & Taylor, J. (2017). Unifying approach to selective inference with applications to cross-validation. arXiv preprint arXiv:1703.06559.

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417–473.

Olive, D. J. (2017). Prediction and statistical learning. Unpublished notes available from http://lagrange.math.siu.edu/Olive/slearnbk.htm.

Rathnayake, R. C. & Olive, D. J. (2019). Bootstrapping some glms and survival regression models after variable selection. Unpublished manuscript at http://lagrange.math.siu.edu/Olive/ppbootglm.pdf.

Rhee, S. -Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., & Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research*, 31, 298–303.

Rinaldo, A., Wasserman, L., G’Sell, M., Lei, J., & Tibshirani, R. J. (2016). Bootstrapping and sample splitting for high-dimensional, assumption-free inference. arXiv preprint arXiv:1611.05401.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., & Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics*, 46, 1255–1287.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111, 600–620.

Tibshirani, R. J. & Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, 822–829.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55, 2183–2202.

Watagoda, L. C. R. P. & Olive, D. J. (2019). Bootstrapping multiple linear regression after variable selection. *Statistical Papers*, 1–20.

Yu, Y., Bradic, J., & Samworth, R. J. (2018). Confidence intervals for high-dimensional Cox models. arXiv preprint arXiv:1803.01150.

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

Zhu, Y. & Bradic, J. (2017). Breaking the curse of dimensionality in regression. arXiv preprint arXiv:1708.00430.

DOI: 10.1002/cjs
APPENDIX 1: OUTLINE OF PROOFS

Proof Outline of Lemma 1
The proof of Lemma 1 consists of three steps:

1. Under the assumptions in Lemma 1, \( \sqrt{n}(\hat{Q} - \text{Err})\) and \( \sqrt{n}(\hat{Q}' - \text{Err})\) converge in distribution to two normal vectors \( Z^\alpha \) and \( Z'^\alpha \) respectively, where \( Z^\alpha \) and \( Z'^\alpha \) have mean 0 and are independent of each other.

2. We show that under assumptions in Lemma 1, we also have
\[
\sqrt{n} \left| \mathbb{E}[\hat{\text{Err}}] - \sum_{k=1}^{m} \text{Err}_k P(O_k) \right| \to \left| \mathbb{E} \left[ \sum_{k=1}^{K} Z_k^\alpha Z_k'^\alpha \right] \right| \cdot \sqrt{n}(\hat{Q}' - \text{Err})_{\mathcal{O}k}, \forall k=1, \ldots, m
\]

3. By the independence between \( Z'^\alpha \) and \( Z^\alpha \), we know the right-hand-side of the above equation is 0.

The same arguments hold for \( \hat{\text{Err}}^j \).

Proof Outline of Lemma 2
The proof of Lemma 2 consists of three steps:

1. Since \( \hat{\beta} \) is the minimizer to the empirical loss with penalty \( \lambda \), by definition we have
\[
\frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda |\beta| \geq \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta})^2 + \lambda |\hat{\beta}|.
\]

2. Since \( \beta \) is the minimizer to the expected loss with penalty \( \lambda \), by the KKT conditions, we know
\[
\mathbb{E}[x(y - x^T \beta)] = \lambda z \quad \text{where} \quad z_j := \frac{\partial |\beta_j|}{\partial \beta_j} \quad \text{is the sub-gradient}. \quad \text{Let} \quad w_i = (y_i - x_i^T \beta)^2 - \mathbb{E}(y - x^T \beta)^2 \quad \text{and} \quad \hat{w}_i = (y_i - x_i^T \hat{\beta})^2 - \mathbb{E}(y - x^T \hat{\beta})^2, \text{then, the inequality from the previous step combined with the KKT condition leads to the following inequality}
\[
\frac{1}{2} (\hat{\beta} - \beta)^T (\mathbb{E}xx^T)(\hat{\beta} - \beta) \leq \frac{1}{2n} \sum_{i=1}^{n} (w_i - \hat{w}_i).
\]

3. We can then prove our statement using basic concentration results for \( (\hat{w}_i - w_i) \), combined with the assumption that \( ||\hat{\beta}||_{\infty} \leq C \) for a constant \( C \).

Proof Outline of Lemma 3
We prove Lemma 3 using straightforward calculus:

1. For any two indexes \( k_t \) with \( t = 1, 2 \), let \( w_i^t \) and \( \hat{w}_i^t \) be the prediction error for model \( k_t \) at sample \( i \) using sample coefficients \( \hat{\beta} \) and population level coefficients \( \beta \) respectively, for \( i = 1, \ldots, n \). By the law of large number and our assumptions, we have
\[
\frac{\sum_{i=1}^{n} (w_i^1 - \frac{\sum_{i=1}^{n} w_i^1}{n})(w_i^2 - \frac{\sum_{i=1}^{n} w_i^2}{n})}{n} \to \Sigma_{k_1, k_2},
\]
and

\[ \mathbb{E} \left( \frac{\sum_{i=1}^{n}(w_i^1 - \sum_{i=1}^{n} \frac{w_i^1}{n})(w_i^2 - \sum_{i=1}^{n} \frac{w_i^2}{n})}{n} \right) \leq C, \]

where \( C \) is a large enough constant.

(2) We then show that

\[ \mathbb{E} \left[ \sum_{i=1}^{n}(\hat{w}_i - \sum_{i=1}^{n} \frac{\hat{w}_i}{n})(\hat{w}_i^2 - \sum_{i=1}^{n} \frac{\hat{w}_i^2}{n}) \right] \to 0, \]

by applying the Cauchy-Schwartz inequality and showing that \( \mathbb{E} \left( \frac{\sum_{i=1}^{n}(\hat{w}_i^2 - w_i^2)^2}{n} \right) \to 0 \), as a result of the consistency and subexponentiality assumptions.

Proof Outline of Theorem 1

We show that \( \sqrt{n}(\mathbb{E}[\hat{\text{Err}}] - \text{Err}^R) \to 0 \) as \( n \to \infty \) by showing that both \( \mathbb{E}[\hat{\text{Err}}] \) and \( \text{Err}^R \) are close to \( \sum_{k=1}^{m} \mu_k P(O_k) \), where \( \mu_k = \mathbb{E}(y - x^T \beta(\lambda_k))^2 \).

1. To show that \( |\sqrt{n}(\mathbb{E}[\hat{\text{Err}}] - \sum_{k=1}^{m} \mu_k P(O_k))| \to 0 \), we bound the left-hand-side by three parts:

   (1) \( I_{11} = |\sqrt{n} \sum_{k \in \mathcal{K}_1} (\hat{Q}_k - \mu_k)I_{O_k} | \), where \( I_{O_k} = \{ \hat{Q}_k \leq \tilde{Q}_k \forall k \in \mathcal{K}_1 \} \) (for \( k \in \mathcal{K}_1 \)) is the selection event that considers only original models that are not too bad.

   (2) \( I_{21} = \sqrt{n} \sum_{k \in \mathcal{K}_1} \mathbb{E}(\hat{Q}_k^\frac{1}{2} - \mu_k)^2 P(O_k \setminus O_k) \) accounts for the difference in the selection events \( O_k \) and \( O_k^1 \) for \( k \in \mathcal{K}_1 \).

2. We can show that \( \sqrt{n}(\mathbb{E}[\hat{\text{Err}}^R] - \text{Err}^R_{\mathcal{K}_1}) \to N(0, \Sigma_{\mathcal{K}_1, \mathcal{K}_1}) \), then by Lemma 1, we have \( I_{1,1} \to 0 \). We can also show \( I_{2,1} \to 0 \) because \( P(O_k^1 \setminus O_k) \to 0 \) applying basic concentration results. For \( I_{3,1} \), use the concentration results, we can show that \( P(O_k) \) goes to 0 faster than \( \mathbb{E}(\tilde{Q}_k^\frac{1}{2} - \mu_k)^2 \) for \( k \not\in \mathcal{K}_1 \), hence, \( I_{3,1} \to 0 \).

3. We then show that \( \sqrt{n}(\text{Err}^R - \sum_{k=1}^{m} \mu_k P(O_k)) \) by showing that

\[ n \mathbb{E} \left( \mathbb{E} \left( |y - x^T \hat{\beta}(\lambda_k)|^2 | \hat{\beta}(\lambda_k) \right) - \mu_k \right)^2 P(O_k) \to 0. \]

and using the Cauchy-Schwartz inequality.

For the second half of the statement, \( \frac{\sqrt{n}(\mathbb{E}[\hat{\text{Err}}^R] - \text{Err}^R_{\mathcal{K}_1})}{1 + \sqrt{n} \Delta} \to 0 \), the proof is similar. However, we have a slower rate because we cannot show that the contribution from the bad models is diminishing (since the selection is not based on the alternative model sequence). Luckily, we can still have that in the worst case scenario, we have a bias term of order \( o\left(\frac{1}{\sqrt{n}} + \Delta\right) \), where \( \Delta = |\text{Err}^R - \text{Err}^R| \) is the underlying true signal.
APPENDIX 2: SIMULATION RESULTS FOR THE BOOTSTRAP $p$-VALUE

In this simulation, we examine the accuracy of the proposed bootstrap $p$-value and compare it to a naive bootstrap with the unadjusted errors. Each column of $X$ represents errors of a model with $n$ samples. We let $Q_j = \frac{1}{n} \sum_{i=1}^n X_{i,j}$ and construct the randomized error $Q_j^b$ and $Q_j^r$ as described in Section 2.1. We let $n = 100$.

Suppose that we have observations $X_{i,j} \sim \mathcal{N}(\mu_j, 1), \ j = 1, 2, \ldots, m, \ i = 1, 2, \ldots, n$. Let $m = 5, 20$. For each $n$, we consider two cases for the underlying $\mu_j$: (1) $\mu_j = 0, \ \forall j = 1, 2, \ldots, m$, (2) $\mu_j \sim \mathcal{N}(0, \frac{1}{n}) \ \forall j = 1, 2, \ldots, m$. For each set of parameter, we repeat 10,000 times the following steps:

1. Construct the de-biased estimate $\hat{Q}^a$ as described in Section 2.1, and the observed error estimate $\hat{Q} = \frac{1}{n} \sum_{h=1}^n Q_{k_h}$, where $k_h$ is the chosen index at $h$th round in the de-biased error estimate algorithm.
2. Bootstrap $B = 1,000$ times, with or without mean rescaling. At each repetition $b$, let $\hat{Q}_b$ and $\hat{Q}_b^r$ be the bootstrap version of the mean error and mean de-biased error, and the bootstrap differences are

$$s_{1,b} = \hat{Q}_b - \hat{Q}, \ s_{2,b} = \hat{Q}_b^r - \hat{Q},$$

where $\hat{Q}$ is the bootstrap population mean across repetitions.
3. Let $\hat{\mu}$ be the true population mean marginalized over the given selection criterion. The $p$-values using the unadjusted error and the de-biased errors are given by:

$$p_{1,b} = \frac{\sum_{b=1}^B 1_{\{\hat{Q}_b - \hat{\mu} \geq s_{1,b}\}}}{B}, \ p_{1,r} = \frac{\sum_{b=1}^B 1_{\{\hat{Q}_b^r - \hat{\mu} \leq s_{1,b}\}}}{B}.$$

$$p_{2,b} = \frac{\sum_{b=1}^B 1_{\{\hat{Q}_b^a - \hat{\mu} \geq s_{2,b}\}}}{B}, \ p_{2,r} = \frac{\sum_{b=1}^B 1_{\{\hat{Q}_b^a - \hat{\mu} \leq s_{2,b}\}}}{B}.$$

Here, $p_{1,b}$ and $p_{2,b}$ are the probability of the bootstrap differences between the estimate and truth being smaller than the difference between our current estimate and the underlying truth, if $p_{1} < \alpha$, it means that the truth is left to the constructed confidence interval at level $\alpha$; similarly, if $p_{2} < \alpha$, it means that the truth is right to the constructed confidence interval.

In this simulation, we know that the test statistics are not degenerate, so we let $\gamma_2 = 0$ in the bootstrap algorithm and we know that the covariance is not degenerate, so we let $\gamma_1 = 0$ in the de-biased error estimate algorithm. In Figure A1, the left and right halves show the empirical CDF plot of the four $p$-values after mean rescaling and without mean rescaling across four parameter settings. The de-biased estimate bootstrap has $p$-value distribution closer to uniform—it has less dependence on the correct underlying distribution than the native bootstrap procedure. Also, the mean rescaling approach leads to better $p$-value distribution.

APPENDIX 3: A POST-SELECTION INFECTION APPROACH CONDITIONAL ON THE SELECTED PENALTY

While the bootstrap $p$-value described in Section 2 tests for a quantity marginalizing out all randomness, the post-selection inference approach conditional on the selected penalty $\lambda_{k^*}$. It corresponds to the case where in practice, we will fix the selected penalty from now on. We consider the following hypothesis
Figure A1: p-Value distribution after 10,000 repetitions. The left plot is the distribution of four p-values after variance rescaling and the right one shows the p-values distribution without variance rescaling. The light/dark blue curves show the c.d.f. of the p-values testing for the observed value from a distribution with mean greater/smaller than the true mean using the unadjusted observation; the red/black curves show the c.d.f. of the p-values using the de-biased estimate.

$$H_0 : \text{Err}_{k^*}^l \leq \text{Err}_{k^*}, \quad \text{vs.} \quad H_1 : \text{Err}_{k^*}^l > \text{Err}_{k^*}. \quad (G_2)$$

In Markovic, Xia & Taylor (2017), let $\lambda_{k^*}$ be the selected penalty, the author have conditioned on (1) the feature set $E$ is selected with penalty $\lambda_{k^*}$ and all predictors, and (2) $\lambda_{k^*}$ is the penalty which minimizes the randomized CV curve, defined as

$$\tilde{Q}_k = Q_k + \frac{\epsilon_k}{\sqrt{n}}, \quad \forall k = 1, 2, \ldots, m.$$ 

where $\epsilon_k \sim \mathcal{N}(0, \tau^2)$ with $\tau^2$ being a constant. For all $\forall j \in E$, the authors then compare the performance between models fitted using OLS with feature set $E$ and $E \setminus j$, trained with all data. In this section, we modify their procedure in the following three aspects: (1) For the model excluding the predictor $j$, we train it with all predictors except for $j$ instead of restricting ourselves to $E \setminus j$. (2) We do not run OLS with all data because we only care about the out-of-sample performance for models produced. (3) We neglect the selection event $A_2$ that $j \in S_{k^*}$ because we want to show that conditioning on the selected penalty $\lambda_{k^*}$ will lead to loss of power.

We look at the test statistics $T = Q_{k^*}^l - Q_{k^*} + \frac{\epsilon}{\sqrt{n}}$, where $\epsilon \sim \mathcal{N}(0, \tau^2)$. Let $\tilde{Q} = (\tilde{Q}_1, \tilde{Q}_2, \ldots, \tilde{Q}_m)^T$. The event $A_1$ can be characterized by $H_{A_1} = \{ \tilde{Q} \in R^m : B_Q \tilde{Q} \leq 0 \}$, where $B_Q$ is the $m \times m$ matrix with a 1 and $-1$ at entry $(k, k)$ and $(k^*, k)$ for $k \neq k^*$, and 0 at other entries. Let the $\Sigma$ be the covariance structure of the vector $\sqrt{n}(T, \bar{Q})^T$. We use $\Sigma_{TT}$ for $n\text{Cov}(T, T)$ and $\Sigma_{T\bar{Q}}$ for $n\text{Cov}(T, \bar{Q})$, etc.

Let $\tilde{Q} = a_\Omega T + N_{\tilde{Q}}$ where $a_\tilde{Q} := \Sigma_{\tilde{Q}T}\Sigma_{TT}^{-1}$. The intuition is that if the three variables $\sqrt{n}T$, and $\sqrt{n}\bar{Q}$ are jointly asymptotically normal, then $N_{\tilde{Q}}$ is asymptotically independent of $T$. We can then condition on $N_{\tilde{Q}}$ and write the constraints in terms of $T$’s asymptotic behaviour and achieve an asymptotic guarantee for the Type I error control.

DOI: 10.1002/cjs
Proposition A1. (Markovic, Xia & Taylor (2017), Theorem 1) Let $T$ be the test statistics. If the following two assumptions hold

1. The selection event $A$ can be characterized in terms of affine constraints over some data vector $D \in S_D = \{ D' | BD' \leq b \}$.
2. The asymptotic joint normality of $(T, D)$ with invertible covariance matrix holds pre-selection $(T, D) \overset{d}{\rightarrow} \mathcal{N} \left( (\theta, \gamma), \begin{pmatrix} \Sigma_{TT} & \Sigma_{TD} \\ \Sigma_{DT} & \Sigma_{DD} \end{pmatrix} \right)$.

Let $D = \Sigma_D D + \Sigma_T T + N_D, (Z_T, Z_D) be the normal vectors from the limiting distribution, then we have

$$P_{\theta, D \in S_D} (\|Z_T - \theta\|_2 \leq \|T - \theta\|_2 | Z_D \in S_D, Z_D - \Sigma_D T, Z_T = N_D) \overset{d}{\rightarrow} \text{Unif}[0, 1].$$

The proposition also works for the one-sided test. By Lemma A1, we know $T$ and $\tilde{Q}$ are asymptotically jointly normal with invertible covariance matrix. Hence, we can construct the $p$-value for any hypothesis value $\theta$ we are interested in based on Lemma A1:

Lemma A1. Let $\theta$ be the hypothesized mean of $T$ and $\sqrt{n}(Z_T - \theta)$ be the normal variable from the limiting distribution of $\sqrt{n}(T - \theta)$. Under the consistency, subexponentiality, and dimension assumptions, we have

1. The following construction of $p$-value achieving the asymptotic uniformity under the simple null hypothesis parameter $\theta$,

$$p_\theta := P_{\theta, \tilde{Q} \in S_{A_1}, D \in S_{A_2}, N_D, N_{\tilde{Q}}}(Z_T \geq T | Z_T \in (a, b)) \overset{d}{\rightarrow} \text{Unif}[0, 1].$$

2. The type I error for the null can be controlled by controlling type I error at $\theta = 0$:

$$\lim_{n \to \infty} P_{H_0}(p_0 \leq \alpha) \leq \alpha.$$

By convention, we let $v_k$ represent the $k^{th}$ element of vector $v$, the maximum of an empty set is $-\infty$ and the minimum of an empty set is $\infty$.

The second part of Lemma A1 is a result of Proposition A2.

Proposition A2. (Lee et al., 2016, Lemma A.1) Let $F_{\theta}(x) := F_{\theta, \sigma^2}(x)$ denote the cumulative distribution function of a Gaussian random variable with mean $\theta$ and variance $\sigma^2$ whose domain of $x$ is $[a, b]$. $F_{\theta}(x)$ is monotone decreasing in $\theta$. 

Received 08 August 2018
Accepted 01 October 2019

The Canadian Journal of Statistics / La revue canadienne de statistique
DOI: 10.1002/cjs