The Canadian Journal of Statistics
Vol. xx, No. yy, 202?, Pages 1–25
La revue canadienne de statistique

Model selection properties of forward selection and sequential cross-validation for high-dimensional regression

Jerzy Wieczorek[1]* and Jing Lei[2]

[1] *Department of Mathematics & Statistics, Colby College, Waterville, ME, U.S.A.*

[2] *Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, U.S.A.*

Key words and phrases: Linear regression; model selection consistency; wrapper forward search.

MSC 2010: Primary 62G09; secondary 62J05

Abstract: Forward selection (FS) is a popular variable selection method for linear regression. But theoretical understanding of FS with a diverging number of covariates is still limited. We derive sufficient conditions for FS to attain model selection consistency. Our conditions are similar to those for orthogonal matching pursuit, but are obtained using a different argument. When the true model size is unknown, we derive sufficient conditions for model selection consistency of FS with a data-driven stopping rule, based on a sequential variant of cross-validation (CV). As a by-product of our proofs, we also have a sharp (sufficient and almost necessary) condition for model selection consistency of "wrapper" forward search for linear regression. We illustrate intuition and demonstrate performance of our methods using simulation studies and real datasets.
The Canadian Journal of Statistics xx: 1–25; 202? © 202? Statistical Society of Canada

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–25; 202? © 202? Société statistique du Canada

1. INTRODUCTION

In the linear regression problem we assume iid data $(\mathbf{X}_i, Y_i)_{i=1}^n$ satisfying

$$Y_i = \mathbf{X}_i^T \beta + \epsilon_i$$

where $\beta \in \mathbb{R}^p$ is unknown, with $\mathbf{X}_i \in \mathbb{R}^p$, and ϵ_i independent noise with mean 0 and variance σ^2. The model selection problem is concerned with finding the support of β: $J_* = \{1 \leq j \leq p : \beta_j \neq 0\}$.

One of the most commonly used model selection procedures is forward selection (FS; Efroymson, 1960). To select the next variable to enter, FS finds the one additional predictor that will minimize the residual sum of squares (RSS). The theoretical properties of FS are still not thoroughly understood, as research has partly focused instead on orthogonal matching pursuit (OMP; Pati et al., 1993), an approximation to FS that has been simpler to study analytically. Because FS has been a mainstay of regression textbooks since at least the time of Draper and Smith (1966) through today (James et al., 2013; Cannon et al., 2019), the operating characteristics of FS deserve to be studied in their own right, even—or perhaps especially—if they turn out to be poor.

* *Author to whom correspondence may be addressed.*

E-mail: jawieczo@colby.edu

In this paper we first derive sufficient conditions under which FS will select the correct subset J_* of active variables after $k = |J_*|$ steps, if k were known. Roughly speaking, assume that the columns of $\mathbf{X}$ have unit variance and absolute pair-wise correlations no larger than $1/(2k-1)$, and that $|\beta_{min}|$, the minimum absolute value of non-zero entries of β, is not too small. Then

$$P(\hat{J}_k = J_*) \to 1\,,$$

where $\hat{J}_k$ is the subset of variables selected by FS after step k. This result is similar to those established for OMP (Tropp, 2004; Cai and Wang, 2011).

Our second contribution is insight into the feasibility of using sample-splitting and cross-validation (CV) as a data-driven stopping rule for FS. While CV is a popular basis for stopping rules, it has its own tuning parameters which are often chosen simply by tradition. We derive specific tuning parameters that guarantee model selection consistency for FS by using a sequential cross-validation (SeqCV) variant. Consider splitting the dataset at random into two parts: a training or construction set of size n_c, and a testing or validation set of size $n_v = n - n_c$. In standard cross-validation (FullCV), we would fit the entire FS model path $\left\{\hat{J}_t : 1 \leq t \leq \min\{n_c, p\}\right\}$ to the training set, then choose $\hat{k}$ to be the value of t whose $\hat{J}_t$ minimizes RSS on the test set. However, this minimum often occurs within a plateau of similar test RSS values after all relevant variables have been chosen, so it is common for FullCV to overfit. By contrast, in SeqCV we choose $\hat{k}$ to be the smallest t whose $\hat{J}_t$ is a local minimizer of test RSS. In either case, the FS path is refit on all n observations up to step $\hat{k}$ to find the final model.

SeqCV has two advantages over FullCV at large sample sizes. First, by alternating the training and test steps, the full path need not be computed and this sequential search for $\hat{k}$ can be much more efficient than FullCV when $k \ll \min\{n_c, p\}$. Second, SeqCV avoids FullCV's tendency to overfit. If the conditions hold for the known-k case above, and both n_c and n_v both grow quickly enough while the training/testing ratio n_c/n_v goes to 0 quickly enough, then

$$P(\hat{J}_{\widehat{k}} = J_*) \to 1 \, .$$

Shao (1997) derives conditions for model selection consistency of cross-validation on a pre-specified set of candidate linear models when k is fixed. However, he prevents overfitting by bounding the number and size of candidate models that are supersets of J_*. Shao's conditions do not hold when this set of overfitting candidate models grows too quickly, as it does in our high-dimensional, random-path setting, which requires a different argument. Our selection consistency proof for SeqCV does not rely specifically on the properties of FS, and we expect a similar result to hold for any path-consistent linear model selection algorithm.

Furthermore, we prove that our sufficient conditions for forward selection with sequential cross-validation (FS+SeqCV) are sharp for a related technique known as wrapper forward search (WrapperFS; Kohavi and John, 1997) applied to linear regression. This is the first known selection-consistency result for a wrapper method. We also discuss the challenging task of selecting n_c/n_v for a given finite dataset and we recommend a rule of thumb at the end of Section 4.

We study model selection consistency: the property that a method selects the correct subset J_* of active variables with probability going to 1. This is a distinct property from prediction consistency, which tells us how quickly a method converges to estimating the regression function. Unfortunately, Yang (2005) shows that these two goals are contradictory. Prediction-optimal methods allow a small although not severe amount of overfitting in sparse settings, as FullCV does. Our SeqCV method's early stopping helps it to avoid overfitting and be selection-consistent, but this also allows occasional underfitting and prevents optimal prediction.

Like SeqCV, the extended BIC (EBIC) of Luo and Chen (2013) is another stopping rule which typically overfits less but underfits more than FullCV. In high-dimensional settings, Luo and Chen (2014) show that EBIC can be a selection-consistent stopping rule when combined with a sequential lasso procedure that is very close to OMP. We believe a similar result for EBIC could be extended to FS, given the similarity between FS and OMP. Nonetheless, the popularity of both FS and CV ensures that their selection properties are still worth understanding.

Other recent work studies the screening property $P(J_* \subseteq \hat{J}_{\widehat{k}}) \to 1$ for FS (Wang, 2009; Charkhi and Claeskens, 2018; Pelawa Watagoda and Olive, 2019).

Feng and Yu (2019) study a related “restricted model selection consistency” property for another CV variant, which they show to perform favourably compared to EBIC in high dimensions. However, both of these properties can allow some spurious variables to enter the model along with all of the true variables, making them weaker than the model selection consistency property we study.

Notation and definitions Subscripts j or h generally refer to a single predictor variable, while J or $\hat{J}$ are index sets for the columns in a particular model: $\hat{J}_t$ is the selected model at step t and $J_* \equiv \{1, \ldots, k\}$ is the true model. In Sections 3.2 and 3.3, $J_h \equiv J_* \cup \{h\}$ is the overfitting model for some $h \notin J_*$, while in Section 4 $J_t \equiv \{1, \ldots, t\}$ is step t of the prespecified model path. $|J|$ is the cardinality of set J. We use bold for multiple columns of the design matrix ($\mathbf{X}$ for the full matrix; $\mathbf{X}_i$ for row i; $\mathbf{X}_J$ for the columns indexed by set J), non-bold for column vectors (X_j for column j; Y for the response; β for the coefficients), and X_{ij} for the element in row i, column j. $\overline{\mathbf{X}}$ is the vector formed by taking the sample mean within each X_j. In the context of split data, $\hat{\beta}$ is always estimated on the training subset. Let $S = n^{-1} \left(\mathbf{X} - \overline{\mathbf{X}}\right)^T \left(\mathbf{X} - \overline{\mathbf{X}}\right)$ denote the sample covariance matrix, and C denote the corresponding sample correlation matrix, with entries $C_{j\ell} \equiv S_{j\ell}/\sqrt{S_{jj}S_{\ell\ell}}$. We use vector notation for inner products and norms: $\langle a, b\rangle = a^T b$ and $\|a\|^2 = \langle a, a\rangle$.

2. BACKGROUND

2.1. The forward selection algorithm

To select the next variable to enter, FS finds the additional predictor that will minimize the RSS. At step t, let $\hat{J}_t$ be the index set of predictors already selected up to this step, with $\hat{J}_0 = \emptyset$. Assume we have centered and scaled Y and all columns of $\mathbf{X}$. Let $Res(Y|X_{\hat{J}_t})$ be the residuals of the response Y on the chosen predictors $X_{\hat{J}_t}$. Then

$$\hat{j}_{t,FS} = \underset{j \notin \hat{J}_t}{\arg\min} \, \|Res(Y|\mathbf{X}_{\hat{J}_t \cup j})\|^2 = \underset{j \notin \hat{J}_t}{\arg\max} \, \frac{\left|\langle Res(Y|\mathbf{X}_{\hat{J}_t}), Res(X_j|\mathbf{X}_{\hat{J}_t})\rangle\right|}{\|Res(Y|\mathbf{X}_{\hat{J}_t})\| \cdot \|Res(X_j|\mathbf{X}_{\hat{J}_t})\|} .$$

The FS algorithm sets $\hat{J}_{t+1} = \hat{J}_t \cup \hat{j}_{t,FS}$ and repeats, until the model size reaches a preset threshold or some other stopping rule is met.

OMP approximates FS by merely finding the predictor most correlated with the current response residuals, as if all predictors were orthogonal:

$$\hat{j}_{t,OMP} = \underset{j \notin \hat{J}_t}{\arg\max} \, \frac{\left|\langle Res(Y|\mathbf{X}_{\hat{J}_t}), X_j\rangle\right|}{\|Res(Y|\mathbf{X}_{\hat{J}_t})\| \cdot \|X_j\|} .$$

The two algorithms will take identical first steps but can differ at any later step.

Appendix S3.1 shows examples where OMP works correctly but FS does not, and vice versa. Since neither method strictly outperforms the other, both deserve study. Yet while there has been recent work on valid statistical inference for FS under standard assumptions of normality (Buja and Brown, 2014; Fithian et al., 2015; Tibshirani et al., 2016), the properties of FS have not been studied as completely as those of OMP and lasso (Tibshirani, 1996). Although statisticians such as Harrell (2015) have justifiably criticized the use of FS on small, noisy datasets, our work provides a much-needed perspective for how FS behaves on modern datasets with massive sample sizes n or dimensions p, without strong distributional assumptions.

2.2. Stopping rules

FS with Sequential Cross-Validation In practice, the performance of FS crucially depends on the stopping rule. In fact, the number of steps taken in FS can be viewed as a regularization parameter (Efron et al., 2004). For sample-splitting, we partition the observations randomly into two sets: a training or construction set s_c of size n_c, and a testing or validation set s_v of size n_v, with $n_c + n_v = n$. Begin to fit the FS model path $\left\{\hat{J}_t : 1 \leq t \leq \min\{n_c, p\}\right\}$ to the training set, and record the estimated coefficient vectors $\{\hat{\beta}_{\hat{J}_t}\}$. After each training step t, estimate the test-set mean squared error (MSE):

$$\widehat{MSE}\left(\hat{J}_t\right) = n_v^{-1} \sum_{i \in s_v} \left(Y_i - \mathbf{X}_i^T \hat{\beta}_{\hat{J}_t}\right)^2 ,$$

and stop at the first model size which is a local minimizer of test MSE:

$$\hat{k}_{Seq} = \min\left\{1 \leq t \leq \min\{n_c, p\} : \widehat{MSE}\left(\hat{J}_t\right) \leq \widehat{MSE}\left(\hat{J}_{t+1}\right)\right\} .$$

Finally, refit the FS model path $\left\{\hat{J}^n_t : 1 \leq t \leq \hat{k}_{Seq}\right\}$ on all n observations and select the model $\hat{J}^n_{\hat{k}_{Seq}}$. We call this stopping rule sequential cross-validation (SeqCV). Besides sample-splitting, the procedure can also be implemented with V-fold CV, where we split the data into V equal-sized subsets, compute a separate

$\widehat{MSE}_\ell\left(\hat{J}_t^\ell\right)$ on each split $\ell \in \{1, \ldots, V\}$, and then choose the final model size based on $\widehat{MSE}(t) = V^{-1}\sum_{\ell=1}^V \widehat{MSE}_\ell\left(\hat{J}_t^\ell\right)$. Similarly, it can be implemented with Monte Carlo CV (MCCV), in which we average the test MSEs over several random splits that all use the same training/testing ratio n_c/n_v.

The large-scale simulations in Section 5.1 and Appendix S3.3 demonstrate FS+SeqCV's promising statistical performance and substantially faster computing time for large datasets as compared to FullCV.

WrapperFS Several of our results in Section 3 also apply to another model selection algorithm, WrapperFS, reportedly commonly used for model selection in the data mining community (Kohavi and John, 1997; Chrysostomou, 2009). With FS+SeqCV, the training path is found using the training data alone, and the test data is only used for the CV-based stopping rule. However, in WrapperFS, the test data is used for both the path selection mechanism and the stopping rule. Our sufficient result for consistency of FS+SeqCV turns out to be sharp for WrapperFS.

During training step t, fit all possible models containing one more variable than in the previous step: $\left\{\hat{J}_{t,j} = \hat{J}_{t-1} \cup j : j \notin \hat{J}_{t-1}\right\}$, and record the estimated coefficient vectors $\{\hat{\beta}_{\hat{J}_{t,j}}\}$. Estimate each corresponding test-set MSE $\widehat{MSE}\left(\hat{J}_{t,j}\right)$ as above. If any of these models improves on the previous MSE, choose it at this step, and otherwise stop:

$$\hat{J}_t = \underset{j \notin \hat{J}_{t-1}}{\arg\min}\, \widehat{MSE}\left(\hat{J}_{t,j}\right) ,$$

$$\hat{k}_{wrap} = \min\left\{1 \leq t \leq \min\{n_c, p\} : \widehat{MSE}\left(\hat{J}_t\right) \leq \widehat{MSE}\left(\hat{J}_{t+1}\right)\right\} .$$

As with SeqCV, we refit the FS model path $\left\{\hat{J}^n_t : 1 \leq t \leq \hat{k}_{wrap}\right\}$ to all n observations and select the model $\hat{J}^n_{\hat{k}_{wrap}}$. Again, the definition above is for sample-splitting and can be extended to V-fold CV or MCCV accordingly.

3. MAIN RESULTS

3.1. Sufficient conditions for path-consistency of FS

We say the random vector $V \in \mathbb{R}^p$ has a sub-Gaussian distribution if there is a constant $c > 0$ such that, for all $u \in \mathbb{R}^p$ with $\|u\|_2 = 1$, we have $\|u^T V\|_{\psi_2} \leq c < \infty$, where for a univariate random variable Z,

$$\|Z\|_\psi = \inf\left\{C > 0 : \mathbb{E}\psi\left(\frac{|Z|}{C}\right) \leq 1\right\} ,$$

and $\psi_2(x) = e^{x^2} - 1$. See van der Vaart and Wellner (1996); van de Geer and Lederer (2013) for some important properties of sub-Gaussian random variables.

Assumption 1 $\mathbf{X}_{n\times p}$ *and* $\epsilon_{n\times 1}$ *are independent random sequences (in* n*) with iid sub-Gaussian rows, with all means 0 and variances* $\mathbb{E}(\mathbf{X}_i\mathbf{X}_i^T) = \Sigma$*, and* $\mathbb{E}(\epsilon_i^2) = \sigma^2$*. For presentation simplicity we further assume that* Σ *is a positive definite correlation matrix (*$\Sigma_{jj} = 1$ *for all* $1 \leq j \leq p$*).*

Throughout Section 3.1 we can weaken Assumption 1 to allow only uncorrelated $\mathbf{X}$ and ϵ. Their independence will be required in Sections 3.2 and 3.3. Also, if Σ is not a correlation matrix, by rescaling $\mathbf{X}$ and β appropriately we can still apply these results to more general covariance matrices, as long as the diagonal entries are uniformly bounded.

The next two assumptions specify the sparsity of β and the relationships between model parameters.

Assumption 2 *The true model satisfies* $|J_*| = \{1, 2, ..., k\}$.

Assumption 3 *As* $n, p, k \to \infty$, *we have*

1. *Bounded coherence, i.e.,* $\mu < (2k-1)^{-1}$, *where* $\mu = \max_{j \neq \ell} |\rho_{j\ell}|$*;*
2. *Signal strength, i.e.,* $|\beta_{min}| \geq c \cdot k\sigma\sqrt{\frac{\log(p)}{n}}$ *for some constant* $c > 0$*;*
3. *Moderate dimensionality, i.e.,* $n^{-1} \cdot \sigma^2 k^2 \log(p) \to 0$.

Remark. Our theory allows p and k to grow with n. In the easier case where p and/or k are fixed, define $N = \max\{n, p\}$. Then in every bound, we can replace factors of the form $\log(p)$ with $\log(N)$.

Theorem 3.1. *Let Assumptions 1, 2, and 3 hold. Then FS is model selection consistent on the path, i.e.,*

$$\mathbb{P}\left(\hat{J}_k = J_*\right) \to 1\,.$$

The proof of Theorem 3.1 is a direct consequence of Proposition S8 and Corollary S2, which are given in Appendix S1 and proven in Appendix S4. Our proof begins with the fixed-noise, fixed-design case and proceeds to random noise and designs.

Previously, others have derived comparable sufficient conditions for correct model selection by OMP. In the noise-free fixed-design setting, Tropp (2004) defines an exact recovery condition (ERC) for OMP, namely $\max_{j>k} \|(\mathbf{X}_{1:k}^T\mathbf{X}_{1:k})^{-1}\mathbf{X}_{1:k}^T X_j\|_1 < 1$, if the columns of $\mathbf{X}$ have unit norm and the first k variables are the true ones. Tropp also shows that the "incoherence condition" $\mu < (2k-1)^{-1}$ implies the ERC.

In the random-noise fixed-design case, Cai and Wang (2011) assume the columns of $\mathbf{X}$ have unit norm and $\epsilon \sim N(0, \sigma^2 I_n)$, and derive model selection consistency of OMP under incoherence and a "beta-min condition," which states that for any $\eta \geq 0$,

$$\min_{i \in 1,\ldots,k} |\beta_i| \geq \frac{2\sigma\sqrt{2(1+\eta)\log p}}{1-(2k-1)\mu} .$$

If we assume the columns of $\mathbf{X}$ have unit variance instead of unit norm, we can replace $|\beta_i|$ with $\sqrt{n}|\beta_i|$ above in order to make the condition comparable with our own Assumptions 1, 2, and 3. However, Cai and Wang's stopping rule depends on the normality of the noise and requires σ to be known.

Besides OMP, similar conditions have been derived for the lasso. Zhao and Yu (2006) and Meinshausen and Bühlmann (2006) independently derived an irrepresentable condition (IC) or neighborhood stability condition, namely that

$$\left|\operatorname{sign}(\beta_{1:k})^T(\mathbf{X}_{1:k}^T\mathbf{X}_{1:k})^{-1}\mathbf{X}_{1:k}^T\mathbf{X}_{(k+1):p}\right| < 1$$

holds element-wise. This IC, which is similar to Tropp's ERC and also implied by $\mu < (2k-1)^{-1}$, is sufficient and "almost necessary" for the lasso's model selection consistency (Bühlmann and van de Geer, 2011).

The incoherence condition $\mu < (2k-1)^{-1}$ is indeed sharp among conditions based only on coherence: there exist cases where the condition is not just sufficient but necessary. To see this, consider the case where $p = k+1$, $\epsilon \equiv 0$ (noise-free), $\beta = (1, ..., 1, 0)^T$, and columns of $\mathbf{X}$ are normalized such that $\mathbf{X}_j^T\mathbf{X}_\ell = -\mu$ if $1 \leq j < \ell \leq k$ and $\mathbf{X}_j^T\mathbf{X}_{k+1} = \mu$ for all $j \leq k$.

Incoherence—which calls for a nearly orthogonal design—is considerably stronger than the ERC or IC, especially at large k. The latter conditions only depend on each spurious predictor's correlation structure with the true variables, while incoherence also restricts all correlations among spurious variables.

It is challenging to find a simple ERC- or IC-like condition for FS, because unlike OMP or lasso, at every step t the FS algorithm rescales each candidate X_j by its residuals on the other predictors already selected in the model. This rescaling causes the FS equivalents of the ERC and beta-min conditions to look different at every step t unless we frame our condition using something much broader such as a coherence bound. Our proofs for FS do lead to the following beta-min condition, derived in Appendix S4.11:

$$\min_{i \in 1,\dots,k} |\beta_i| \geq \frac{c\sigma\sqrt{\log p}}{(R_{min} - M) \cdot \lambda_{min}(\mathbf{X}_{1:k}^T \mathbf{X}_{1:k})} \,. \tag{1}$$

$R_{min} = \min_{j>k, J \subset \{1:k\}} Res(X_j|\mathbf{X}_J), \qquad M = \max_{j>k} \|(\mathbf{X}_{1:k}^T \mathbf{X}_{1:k})^{-1} \mathbf{X}_{1:k}^T X_j\|_1$, and λ_{min} denotes a minimum eigenvalue. However, like ERC and IC, it relies on knowing the true support J_*, even though our goal in model selection is to learn J_*.

Our simulations in Section 5.1 suggest that the stringent sufficient condition based on coherence is not usually necessary unless Σ has a particularly disadvantageous structure. Finally, we can relax the incoherence condition if we assume that Σ is row-sparse.

Corollary 3.1. *Assume that each row of Σ has no more than s nonzero off-diagonal entries for some $0 \leq s < k$. Under Assumptions 1 and 2, and a modification of Assumption 3 replacing $\mu < (2k-1)^{-1}$ by $\mu < (3.4s)^{-1}$, FS is model selection consistent on the path.*

The result follows directly from Proposition S8 with Corollary S5.

On the other hand, even for orthogonal Σ, Su (2018) shows that moderate k can be a problem for stepwise algorithms like FS and lasso. Above a certain extreme-sparsity threshold, larger k causes the first spurious variable to be selected earlier and earlier, making support recovery impossible. Our supplemental simulations in Appendix S3.3 demonstrate the same phenomenon.

3.2. Model selection consistency of FS+SeqCV

Now, assume we do not know the value of k but we estimate it by FS with sequential data-splitting, or FS+SeqCV, as defined in Section 2. This procedure is also model selection consistent, under the additional conditions below.

Assumption 4 *As $n, p, k \to \infty$, we require the following:*

1. *Assumption 3 holds on the training sample, with n replaced by n_c;*
2. *Balanced coefficients, i.e., $\frac{\beta^2_{min}}{\beta^2_{max}} \geq c \cdot \max\left\{k\sqrt{\frac{\log(k)}{n_c}}, \frac{k^2\log(k)/n_v}{\beta^2_{min}/\sigma^2}\right\}$ for some constant $c > 0$;*
3. *Moderate dimensionality and training/testing ratio, i.e., $\max\left\{\frac{k}{n_c}, \frac{n_c}{n_v}\right\} \cdot kp^2\log(p) \to 0$.*

The condition that coefficients be balanced prevents underfitting by ensuring that estimation error in large coefficients does not cause us to stop before the smallest coefficients are selected.

Next, the condition $p^2 \ll n_c$ is much stronger than what we needed for Theorem 3.1, where $p > n$ was possible at every n. For a particular spurious variable h, consider the overfitting model $J_h = J_* \cup h$ and let $B_h \equiv \mathbb{E}_v\left(\widehat{MSE}(J_h) - \widehat{MSE}(J_*)\right)$ be the difference in risks between this incrementally larger model and the true model, as fitted to the observed training data. We say we make a **training mistake** if $B_h < 0$. The condition $n_c \to \infty$ ensures that a given B_h has the correct sign, while the p^2 term comes from a union-bound argument over all h.

Our required rate of $p^2/n_c \to 0$ comes from a careful analysis of the expansion of B_h in which we exploit a cancellation between the $\tilde{\beta}_{J_h}$ and an error term, where $\tilde{\beta}_{J_h}$ is the regression coefficient for the noise regressed on X_h after projecting out $\mathbf{X}_{J_*}$. In contrast, a straightforward argument directly using rates of convergence for $\tilde{\beta}_{J_h}$ would lead to a much stricter requirement of $p^4/n_c \to 0$ after the union bound.

Similarly, the condition $p^2 \ll \frac{n_v}{n_c}$ comes from a union-bound argument applied to the conditions for avoiding a different mistake. We say we make a **model selection mistake** if we observe $\widehat{MSE}(J_h) < \widehat{MSE}(J_*)$ in our combined training and testing samples. A small ratio n_c/n_v can prevent overfitting by ensuring that the difference in risks obtained from the training sample estimates dominates the test-set error in estimating this difference. The same idea appeared in Yang (2007) in finite dimensional problems. Our additional requirement on n_c/n_v reflects the impact of dimensionality.

Theorem 3.2. *Let Assumptions 1, 2, and 4 hold. Then FS+SeqCV is model selection consistent, whether we use the sample-splitting version or MCCV with a fixed number of splits:*

$$\mathbb{P}\left(\hat{J}_{\hat{k}_{Seq}} = J_*\right) \to 1.$$

The proof of Theorem 3.2 is given in Appendix S2.

When using FS to determine the model path, we conjecture in Section 3.3 that Assumption 4.3 could be weakened. Even so, our proof that the probability of a model selection mistake vanishes is of independent interest, since its worst-case setup corresponds to the WrapperFS algorithm defined in Section 2. In Section 3.3, we show that Assumption 4.3 is not only sufficient but also "almost" necessary for WrapperFS—the condition is sharp in the sense that there are situations where the assumption cannot be substantially weakened.

3.3. On the effect of p

Here we consider a very simple case to show that the condition $p^2/n_c \to 0$ in Assumption 4.3 is essentially necessary for selection consistency of WrapperFS.

Assumption 5 *The true model J_* is $Y = \mu + \epsilon$. We compare this against p spurious univariate models: $Y = \beta_{0h} + \beta_{1h}X_h + \epsilon$, for candidate covariate $h \in 1, \ldots, p$.*

Assumption 6 $\mathbf{X}_{n\times p}$ *and $\epsilon_{n\times 1}$ are independent random sequences (in n) with iid Gaussian rows. Each row has mean 0 and variances $\mathbb{V}(\mathbf{X}_i) = I$ and $\mathbb{V}(\epsilon_i) = 1$.*

Assumption 7 *As $n \to \infty$, the number of candidate variables p grows "too quickly"; that is, $\liminf p^2/n_c \geq \Gamma$ for some constant $\Gamma > 0$. It does not matter whether or not the training/testing ratio n_c/n_v goes to 0.*

Theorem 3.3. *Let Assumptions 5, 6, and 7 hold. Then the probability that WrapperFS makes a model selection mistake does not vanish:*

$$\liminf_{n\to\infty} \mathbb{P}\left(\min_h \widehat{MSE}(J_h) < \widehat{MSE}(J_*)\right) \geq 0.16(1 - e^{-\sqrt{\Gamma}/2}) > 0\,.$$

On the other hand, the conditions $p^2 \ll n_c$ and $p^2 \ll n_v/n_c$ do not appear necessary for FS+SeqCV. These conditions arise from a union bound applied to WrapperFS, which uses the test set to evaluate *every* spurious variable and continues if at least one trained slope fits the test data well. Thus, WrapperFS overfits more often at higher p. However, FS+SeqCV will rarely require that same union bound in practice. After all correct variables have been chosen, FS+CV will use the test set to evaluate only *one* spurious variable: the one with the lowest training-data estimate of risk. If $p - k$ is large, the spurious variable that overfits most to the training data is unlikely to fit the test data well, so FS+SeqCV should stop without adding any spurious variables.

The simulations in Figure 1 illustrate this intuition. We draw n sample outcomes from a true null model along with p spurious predictors from an orthogonal random Normal design, as in Assumptions 5 and 6. For the WrapperFS and FS+SeqCV methods, respectively, the top and bottom subplots of Figure 1 contain heatmaps showing the proportion of correct model selection results for different combinations of p and n_c, estimated from 500 replications of each combination. In order to evaluate both conditions of interest using a single pair of figures, we choose $n_c = \sqrt{n}$ so that $\frac{n_v}{n_c} = \sqrt{n} - 1 \approx n_c$. As per Theorem 3.3,

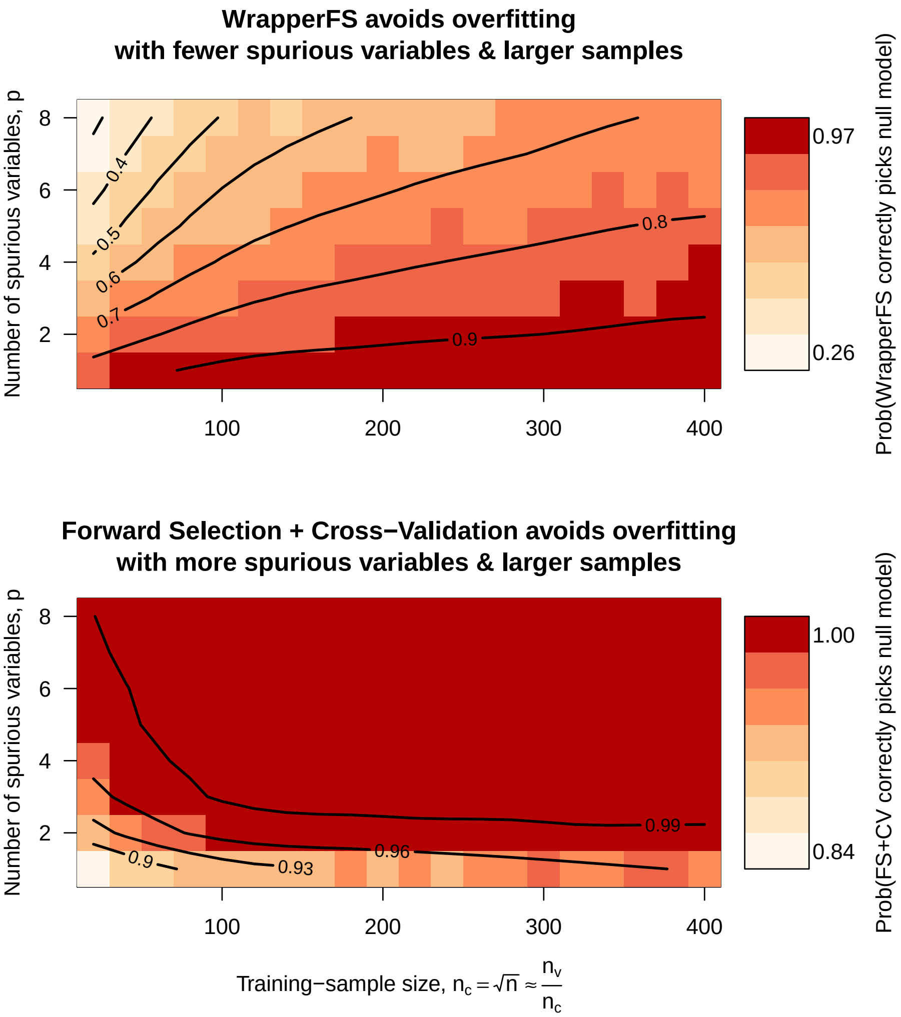

FIGURE 1: Top figure: $\mathbb{P}$(correctly choose null model) for WrapperFS. The contours are generally shaped roughly like $p = \sqrt{n_c} \approx \sqrt{\frac{n_v}{n_c}}$. Bottom figure: For FS+SeqCV, the contours are not at all like $p = \sqrt{n_c} \approx \sqrt{\frac{n_v}{n_c}}$. Contour lines estimated from 2D loess fit, based on 500 simulations in each cell of the $p \times n_c$ grid.

WrapperFS is model selection consistent only if $p^2/n_c \to 0$ and $p^2 \cdot \frac{n_c}{n_v} \to 0$. However, for FS+SeqCV, at every n_c and $\frac{n_v}{n_c}$, the proportion of correct model selection results rises with p, as we conjecture above. We leave this as a focus for future work.

4. PRACTICAL CHOICE OF TRAINING RATIO FOR LARGE SAMPLES

Traditional high training ratios n_c/n, as in 5-fold or 10-fold CV, tend to avoid underfitting models but are prone to overfitting at any n. However, at large n, the chance of underfitting is low at nearly every training ratio, such that it may be safe to reduce n_c/n in order to avoid overfitting as well. We suggest some rules of thumb for making this tradeoff.

In this section, let us assume that n is sufficiently large for FS to select a correct path. In this setting, we build on intermediate results from Zhang (1993), who assumes that the model path is fixed in advance, the true model is indeed on this path, and p and k are fixed as n grows. Under multifold CV (MCV; a scheme in which one averages the test MSEs from all possible splits with the same training ratio), Zhang's Corollary 1 provides an exact asymptotic distribution for the probability of correct model selection, which decreases monotonically as n_c/n and $p - k$ increase. Equivalently, this is the asymptotic probability of avoiding overfitting, since the probability of underfitting goes to 0 regardless of n_c/n.

However, those probabilities are for FullCV. With SeqCV on a correct fixed path, we only need the probability that FS+SeqCV stops at the correct model instead of going one step further. By Zhang's asymptotic probability for avoiding overfitting evaluated at $p - k = 1$,

$$\mathbb{P}\left(\hat{k} > k\right) \approx 1 - \mathbb{P}\left(\chi_1^2 < \left(1 + \frac{n}{n_c}\right)\right) .$$

When $n_c/n \leq 1/10$, this probability of an overfitting mistake becomes negligible with $\mathbb{P}(\hat{k} > k) < .001$. Even with massive n, there is rarely a pragmatic need to reduce the training ratio past $n_c/n = 1/10$ for SeqCV.

Next, in order to relate n and n_c/n to the probability of underfitting, we refine an intermediate step in Zhang's proof of his Theorem 1, which derives relevant expressions for $\widehat{MSE}(J_t)$.

Corollary 4. *Let $\epsilon \sim N(0, \sigma^2)$ and let $\mathbf{X}$ be a deterministic sequence. Assume the model path is fixed and the predictors are ordered, so that model J_t corresponds to using the first t predictors. Consider comparing true model J_* (of size k) against a particular underfitting model J_t (of size t), where $J_t \subsetneq J_*$. For $t < k$, define $b_{J_t} = \liminf_{n\to\infty} n^{-1}(\mathbf{X}\beta)^T P_{J_t}^{\perp}\mathbf{X}\beta > 0$. Under additional technical conditions A', B, C', and D, stated in Appendix S4.12, we can ensure that $\lim_{n\to\infty} \mathbb{P}(\text{correctly choose } J_* \text{ over } J_t) \geq 1 - \alpha$ by choosing a training ratio n_c/n of at least*

$$\frac{n_c}{n} \geq \left(\frac{\left(\sqrt{\frac{b_{J_t}}{\sigma^2/n}} - Z_{1-\frac{\alpha}{2}}\right)^2 + \chi^2_{(k-t),\alpha/2} - Z^2_{1-\frac{\alpha}{2}}}{k} - 1 \right)^{-1}. \qquad (2)$$

Corollary 4 only considers comparing J_* against a single underfitting model J_t. However, since b_{J_t} and $\chi^2_{(k-t),\alpha/2}$ are both monotonically nonincreasing along the model path (as t rises toward k), the worst case should be the model with $t = k - 1$.

Choosing n_c/n We argued above that a training ratio around $n_c/n = 1/10$ is more than adequate to avoid overfitting, whereas 10-fold CV's $n_c/n = 9/10$ is commonly used to avoid underfitting. We might expect Corollary 4 to help us tune n_c/n and balance these competing tendencies. However, Equation 2 shows that this is impractical unless we have fairly good knowledge about b_{J_t} and k.

Unless the tolerated probability of failure α or the signal-to-noise ratio (SNR) are miniscule, the Z and χ^2 terms are negligible for large n. Equation 2 then implies we need $\sqrt{1 + \frac{n}{n_c}} \leq \sqrt{\frac{nb_{J_{k-1}}}{\sigma^2 k}}$, or $\sqrt{1 + \frac{n}{n_c}} \leq \sqrt{\frac{n}{k}} \frac{|\beta_{min}|}{\sigma}$ if Σ is close to orthogonal. Hence, there is only a narrow range of $|\beta_{min}|/\sqrt{k\sigma^2}$ where it makes sense to decide between $n_c/n = 1/10$ and $n_c/n = 9/10$: $\sqrt{1 + \frac{10}{1}}/\sqrt{1 + \frac{10}{9}} \approx 2.28$. The choice of n_c/n (or equivalently the choice of V for V-fold CV) is so sensitive that we need to know $|\beta_{min}|/\sqrt{k\sigma^2}$ to within a factor of 2, which is implausible in many modern high-dimensional regression settings, even with a pilot study.

Instead, we propose the following rule of thumb, whose use we illustrate in Section 5.2: If n is large and the signal is strong, i.e., if we confidently believe $\sqrt{\frac{nb_{J_{k-1}}}{\sigma^2 k}} \gg \sqrt{1 + \frac{10}{1}} \approx 3.32$, then we can safely use a low training ratio of $n_c/n = 1/10$, avoiding both under- and overfitting. For V-fold CV, this means an "inverted" approach to 10-fold CV: create 10 folds, but train on only one at a time and test on the other nine. Otherwise—if n is not large, or our initial guess of $\sqrt{\frac{nb_{J_{k-1}}}{\sigma^2 k}}$ is too small or imprecise—a conventional training ratio such as

$n_c/n = 8/10$ or $9/10$ (standard 5-fold or 10-fold CV) is safer. We will be prone to overfitting but at least ought to avoid underfitting.

Tradeoffs between underfitting and overfitting In Appendix S3.2, we illustrate the above approximations for $\mathbb{P}(\text{underfit})$ and $\mathbb{P}(\text{overfit})$ and compare them to empirical estimates. Figures S1 and S2 confirm that high training ratios avoid underfitting, low ratios avoid overfitting, and there is only a narrow range of $\sqrt{b_{J_{k-1}}}$ in which it makes sense to tune n_c/n.

5. SIMULATION STUDIES AND REAL-DATA EXAMPLES

5.1. Simulation design and results

We study stopping-rule procedures chosen for comparison with 5-fold CV, one of the most common CV variants. We contrast standard V-fold vs. an "inverted" variant designed for small training ratios (training on one fold and testing on the remaining $V-1$ folds). We also contrast our SeqCV vs. standard FullCV. Below we report simulations of the probability of correct model recovery. Additional simulation results for false positives and false negatives, as well as heavy-tailed noise, are reported in Appendix S3.3.

We simulated a range of true model sizes $k \in \{5, 25, 125\}$, dimensions $p \in \{10, 50, 250, 1250\}$, and sample sizes $n \in \{50, 250, 1250, 6250\}$ (omitting the impossible settings where $k > p$ or $k > n$). We found that this range for n, together with a small $\beta_{min} = 0.2$, adequately contrasts the low SNR and high SNR cases. The k nonzero coefficients were drawn from a Uniform distribution, then shifted and scaled to have the range $[\beta_{min} = 0.2, \beta_{max} = 2]$. Noise terms were drawn independently as $\epsilon \sim N(0, 1)$. The design matrix $\mathbf{X}$ was drawn from $N(0, \Sigma)$ with two settings.

In Setting 1, $\Sigma_1(\mu) = (1 - \mu)I + \mu\mathbf{1}\mathbf{1}^T$ is a constant-correlation matrix with all off-diagonal elements set to $\mu \in \{1/(2k), 5/(2k)\}$, allowing us to compare results based on whether the coherence was just below or far above the theoretical threshold of $\mu < (2k - 1)^{-1}$.

In Setting 2, $\Sigma_2(\mu)$ has the following "unfavorable" correlation matrix structure:

$$\Sigma_{j\ell} = \begin{cases} 1 & \text{if } j = \ell\,, \\ -\mu & \text{if } j \neq \ell \text{ and } j, \ell \leq k\,, \\ \mu & \text{if } j \neq \ell \text{ and } j \text{ or } \ell \in k+1, \ldots, p\,. \end{cases}$$

Here, the coherence condition is not only sufficient but also necessary to recover the simple model where $\epsilon = 0$ and $\beta_j = 1$ for $j \leq k$. This structure may not be positive definite for certain combinations of k, p, and μ, but we report results for $k = 10$, $p = 11$, and $\mu \in \{1/19, 1/10\} = \{1/(2k - 1), 1.9/(2k - 1)\}$, where Σ is positive definite.

We run at least 400 replicate simulations at every combination of data-generation settings, independently generating new datasets and running every estimation procedure on that dataset. In Figures 2-3 and Appendix S3.3, error bars show $\pm 2 \cdot SE$ as approximate marginal 95% confidence intervals.

As expected, in each subplot of Figure 2 the model selection problem becomes easier with increasing SNR. Higher μ makes the problem harder but not impossible, as we see by comparing the left and right halves of Figure 2. Higher p and higher k both make the problem harder, causing ever-higher values of n to count as "low SNR" conditions.

Figure 2 also shows results for FS with V-fold CV-based stopping rules, at two training ratios: 5-fold (dashed lines) and inverted 5-fold (dotted lines). We see similar patterns as for oracle FS, but with lower success rates when the stopping rule is only an estimate. As expected, the low training ratio performs the worst at low n, due to substantial chance of underfitting (despite low chance of overfitting). However, the same low training ratio performs best at high n, due to negligible chance of underfitting when SNR is high, while overfitting never becomes negligible for high training ratios even at high SNR. Standard 5-fold CV is better than inverted 5-fold for small and moderate n but its success probability tends to plateau beyond $n = 1250$.

In addition to the effects of high vs. low training ratio, Figure 2 also contrasts between SeqCV vs. FullCV (dark vs. light colors, respectively). In several low-k cases, 5-fold FullCV overfits more than 5-fold SeqCV does at high SNRs, while in a few high-k cases, 5-fold SeqCV stops too early more often than 5-fold FullCV does. However, the differences between SeqCV and FullCV are otherwise negligible.

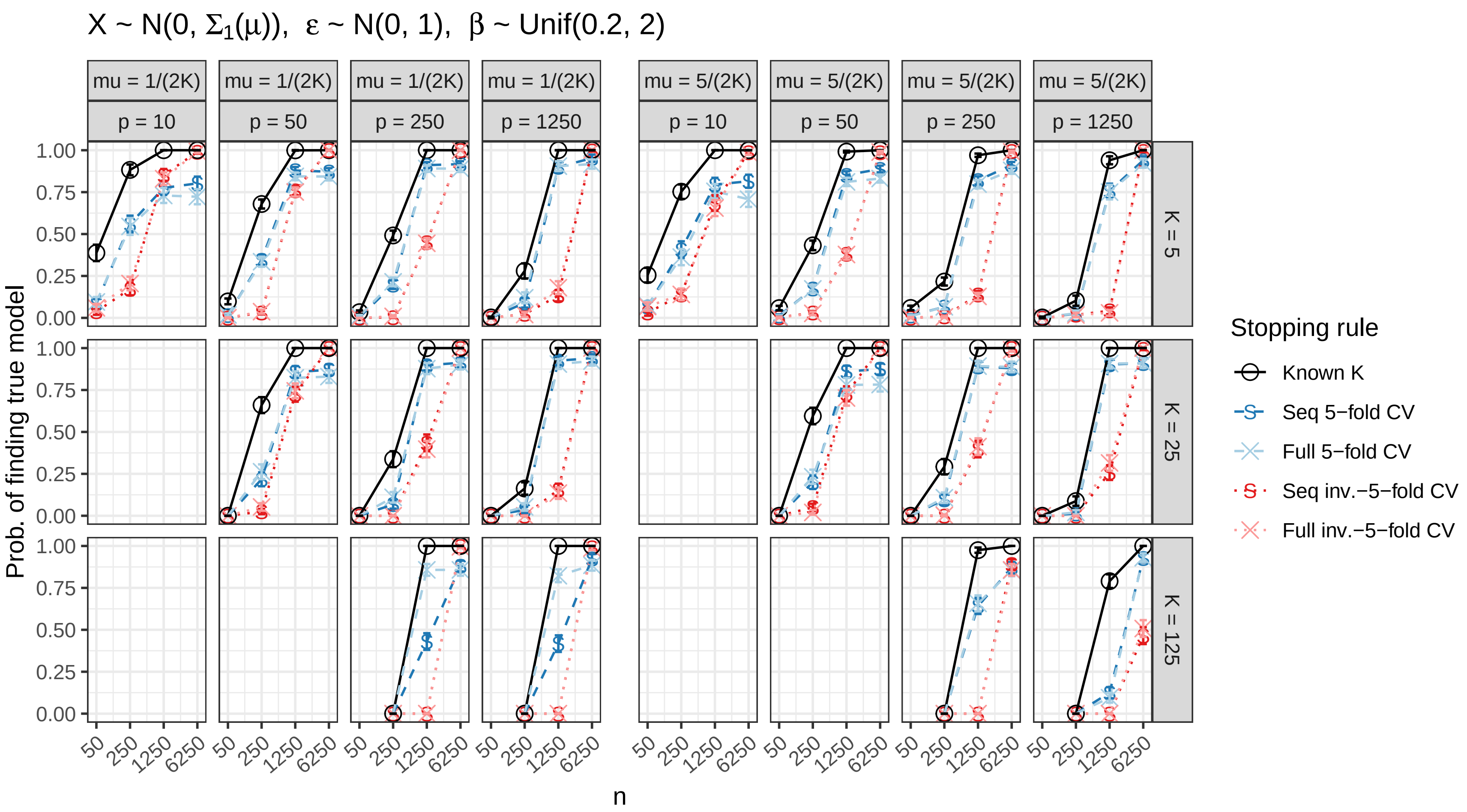

FIGURE 2: Proportion of correct model selections over 400 repetitions with constant-correlation setting $\Sigma_1(\mu)$. The incoherence condition is met in the left half of the figure, but not in the right half. Empty subplots correspond to impossible combinations $k > p$ or $k > n$. Error bars show $\pm 2 \cdot SE$.

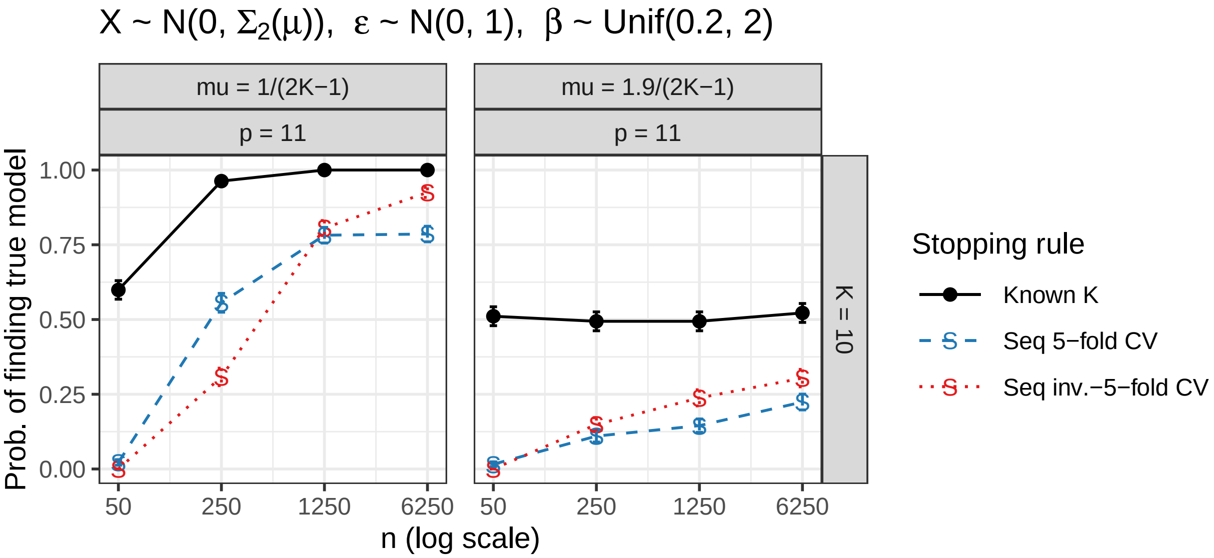

FIGURE 3: Proportion of correct model selections over 1000 repetitions with "unfavorable" correlation setting $\Sigma_2(\mu)$. When incoherence condition is not met (right subplot), the probability of success is bounded away from 1. Error bars show $\pm 2 \cdot SE$.

By contrast, Figure 3 shows the effect of the "worst-case" correlation structure $\Sigma_2(\mu)$. This simulation setup was similar to selected subplots from Figure 2, except with a different Σ. The $\mu < (2k-1)^{-1}$ case is similar in both figures, but the high-μ case is dramatically worse in Figure 3. Here, the success probability is stuck at around 0.5 for oracle FS and even lower for FS+CV, since the structure of $\Sigma_2(\mu)$ is designed to cause a mistake when $\mu \geq (2k-1)^{-1}$. This simulation illustrates that our coherence condition is sharp, even though it may not be necessary under other structures for Σ as seen in Figure 2 and supplemental Figure S5.

5.2. Real-data example: Million Song Dataset

We illustrate FS+SeqCV on a large dataset, where a small training/testing ratio and the early stopping of SeqCV can be expected to improve both run-time and probability of correct model selection. We use the year-prediction problem extract of the Million Song Dataset (MSD) assembled by Bertin-Mahieux et al. (2011). At $n = 515{,}345$ and $p = 90$, this is one of the largest regression datasets currently on the UCI Machine Learning Repository (Lichman, 2017). All of the predictors are continuous and have no missing values.

Other authors have previously used this dataset to illustrate regression methods for large-scale data. Zhang et al. (2015) used the MSD to illustrate a scalable variant of kernel ridge regression (KRR), while Ho and Lin (2012) used the MSD as a test case for linear support vector regression (SVR) vs. kernel SVR.

Task and data description

Our task is to predict the "continuous" release year (between 1922 and 2011) of each song in the dataset, using 90 continuous predictors all based on the acoustic property of "timbre." According to the documentation for The Echo Nest "Analyze" API used to preprocess the MSD (Jehan, 2010), "*timbre* is the quality of a musical note or sound that distinguishes different types of musical instruments . . . and is derived from the shape of a segment's spectro-temporal surface, independently of pitch and loudness."

Each record in the MSD is one song. That song is partitioned into short time segments, and 12 timbre coefficients are computed on each segment to approximate the segment's spectral surface as a linear combination of 12 basis functions. Finally, the 12 averages, 12 variances, and 66 covariances of these coefficients

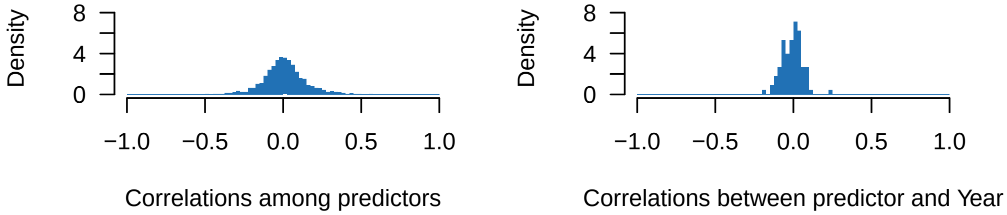

FIGURE 4: Histograms of correlations in the MSD dataset.

(across time segments within a song) are computed to create the 90 features in the MSD.

Relating release year linearly to timbre may not be an ideal scientific model, but we have not found nonlinear methods to be substantially better. A linear regression with all $p = 90$ variables achieves a holdout root mean squared error (RMSE) of 9.51 years, with $R^2 \approx 0.24$. Zhang et al. (2015) report their nonlinear kernel ridge regression achieves holdout pseudo-$R^2 \approx 0.31$, the same as our own best attempt at nonlinear regression using random forests. Hence, linear regression has little room for improvement and appears to be an adequate predictive model for this dataset.

Finally, Figure 4 shows several high correlations between pairs of predictors, but not many. Although we do not meet the coherence threshold used for our theoretical results, we are not too concerned in light of the simulation results in Section 5.1. Also, the correlation matrix of our predictors has condition number 13.3, which is below the conventional multicollinearity cutoff of 30.

Model selection and evaluation

We illustrate the use of our proposed method, FS+SeqCV with a low training/testing ratio, compared against several alternatives. The MSD is published with a pre-determined 90/10 split of $463{,}715$ learning cases and $51{,}630$ holdout cases. We perform CV by splitting the $463{,}715$ learning cases further, and we report performance on the $51{,}630$ holdout cases.

Training an intercept-only model (guessing every song's release year as 1998.4) has a test-data RMSE of 10.85, while the full OLS model has a test-data RMSE of 9.51. This difference in RMSEs translates to 1 year and 4 months, so even the full linear model does not improve predictions dramatically on average. Hence, we merely hope to find a sparser linear model whose holdout RMSE is nearly as good as the full model's, for the usual benefits of model selection: better interpretability, fewer predictors to collect, etc.

Following our heuristic advice in Section 4, we believe that a 10/90 split is reasonable here. The training set has a large $n = 463{,}715$, so we can safely use a training ratio of $n_c/n = 1/10$ if we also believe that $\sqrt{\frac{nb_{J_{k-1}}}{\sigma^2 k}} \gtrsim \sqrt{1 + \frac{10}{1}} \approx$ 3.32. Let us decide that it does not make sense to estimate a sparse model here unless it has at most $k \leq p/3 = 30$ nonzero coefficients. For the full OLS model, $\hat{\sigma} \approx 9.6$, and the top 35 $|\hat{\beta}_j|$ in the full model are all above 0.30, so it seems reasonable to assume $|\beta_{min}| \gtrsim 0.3$. This leads to $\sqrt{\frac{nb_{J_{k-1}}}{\sigma^2 k}} \approx \sqrt{\frac{463{,}715}{30}} \times \frac{0.3}{9.6} \approx$ $3.9 > 3.32$, so it appears reasonably safe to use $n_c/n = 1/10$. However, for the sake of comparison, we also run FS+SeqCV with a high training ratio, as well as FS+FullCV at both training ratios. Finally, we also report results for the null (intercept-only) model and the full OLS model (all 90 predictors).

For the null and full models, we train directly on all $463{,}715$ learning cases and report performance on the $51{,}630$ holdout cases. For the CV-based stopping rules with 10/90 training/testing ratios, we fit a model path on the first $n_c = 46{,}371$ cases as a training set, then use the next $n_v = 417{,}344$ cases as a test set to select one model from that path. We finally refit the chosen model on the full learning set and report its performance on the holdout set. For the 90/10 training/testing ratios, we reverse the roles of these same training and test sets.

For each approach, Table 1 reports the size of the selected model, holdout RMSE estimates (in years), and average model selection computation time (in seconds) over 10 runs on a laptop with an Intel Core i7 CPU and 16GB of RAM. As we move down the rows of Table 1, we modestly reduce RMSE but dramatically increase model size and computation time.

First, we note in Table 1 that a lower training/testing ratio (10/90 vs. 90/10) does not substantially change the selected model's RMSE, but it does tend to choose a sparser model, as we expect for such large-n situations. Second, the SeqCV stopping rule tends to choose a substantially sparser model than FullCV. These sparser models do tend to have slightly higher holdout RMSE, but by no more than 0.1 on the Year scale, which corresponds to 1.2 months—a negligible

TABLE 1: Million Song Dataset selected model sizes, holdout RMSEs, and mean model selection computation times (with 95% MOEs from 10 runs), under different stopping rules.

Stopping rule	$\hat{k}$	RMSE (years)	Time ± MOE (seconds)
Null model	0	10.852	—
FS+SeqCV, 10/90	23	9.598	1.30 ± 0.03
FS+SeqCV, 90/10	29	9.562	3.18 ± 0.04
FS+FullCV, 10/90	60	9.515	3.71 ± 0.11
FS+FullCV, 90/10	76	9.511	4.50 ± 0.06
Full model	90	9.510	—

difference, especially with data recorded to the nearest year. In the sparsest case, FS+SeqCV at the 10/90 training/testing ratio selects a model with 23 variables, around a quarter of the original 90 predictors. This is a considerable reduction in model size with negligible effect on predictive performance.

Table 1 also reports the approximate runtime for each selection algorithm. Using a small 10/90 split can be substantially faster than a large 90/10 split, because the computationally-expensive training is run on far less data. Likewise, using SeqFS can be substantially faster than FullCV, because it is possible to stop quite early without building a full model path up to all 90 variables.

In short, our suggested algorithm selects a model which performs almost identically to the largest model in our scope, but which requires far fewer predictor variables and considerably less computation time than does traditional FullCV with a high training ratio. If we are using a linear model and we have massive n, the combination of SeqCV and low training ratio can improve sparsity and computation speed dramatically with minimal impact on predictive performance. Of course, examples can exist where a small early uptick in estimated CV error causes FS+SeqCV to underfit, but these are most probable in small-data, low-signal settings, where correct model selection is hopeless for any algorithm.

6. DISCUSSION

We have derived and illustrated conditions under which FS is model selection consistent, either assuming that the model size k is known or using a data-driven stopping rule based on SeqCV. Our path-consistency conditions may prove useful for future work on other stopping rules that assume a correct path, such as BIC, or on claims for an estimator's oracle property. We also demonstrate the benefits of using a low CV training ratio under the suitable conditions outlined in our rule of thumb from Section 4.

However, we remind practitioners that if predictors are highly correlated, the FS path may be wrong or SeqCV may underfit badly. Alternative model selection algorithms such as the lasso's less-greedy search may be less likely to choose spurious variables as early as FS. Alternative stopping rules such as EBIC or FullCV may be less likely to underfit than SeqCV, and their degree of overfitting is still unlikely to be severe if the model is sparse. Furthermore, previous authors such as Harrell (2015) have argued that automatic regression model selection is rarely a good idea, as the choice of model is noisy and much less interpretable than it appears. We agree with Harrell in the noisy small-sample setting where FS has traditionally been applied. As our theorems and simulations demonstrate, there are no guarantees that FS will do a good job of selection when we have low signal, high noise, small samples, and high correlations.

ACKNOWLEDGEMENTS

We thank the Referees and the Review Editor for their helpful comments. Our research was partially supported by National Science Foundation grant DMS-1553884.

BIBLIOGRAPHY

Bertin-Mahieux, T., D. P. Ellis, B. Whitman, and P. Lamere (2011). The Million Song Dataset. In *ISMIR*, Volume 2, pp. 10.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Buja, A. and L. Brown (2014). Discussion: “A significance test for the lasso”. *The Annals of Statistics 42*(2), 509–517.

Cai, T. T. and L. Wang (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory 57*(7), 4680–4688.

Cannon, A. R., G. W. Cobb, B. A. Hartlaud, J. M. Legler, R. H. Lock, T. L. Moore, A. J. Rossman, and J. A. Witmer (2019). *STAT2: Modeling with Regression and ANOVA* (2nd ed.). W.H. Freeman.

Charkhi, A. and G. Claeskens (2018). Asymptotic post-selection inference for the Akaike Information Criterion. *Biometrika 105*(3), 645–664.

Chrysostomou, K. (2009). Wrapper feature selection. In *Encyclopedia of Data Warehousing and Mining, Second Edition*, pp. 2103–2108. IGI Global.

Draper, N. R. and H. Smith (1966). *Applied Regression Analysis* (1st ed.). John Wiley & Sons, Inc.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers*, 191–203.

Feng, Y. and Y. Yu (2019). The restricted consistency property of leave-n_v-out cross-validation for high-dimensional variable selection. *Statistica Sinica 29*(3), 1607–1630.

Fithian, W., J. Taylor, R. Tibshirani, and R. Tibshirani (2015). Selective sequential model selection. *arXiv preprint arXiv:1512.02565*.

Harrell, F. (2015). *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal regression, and Survival Analysis*. Springer.

Ho, C.-H. and C.-J. Lin (2012). Large-scale linear support vector regression. *Journal of Machine Learning Research 13*(Nov), 3323–3348.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning*. Springer.

Jehan, T. (2010). *Analyze Documentation*. Echo Nest Analyze API version 2.2.

Kohavi, R. and G. H. John (1997). Wrappers for feature subset selection. *Artificial Intelligence 97*(1-2), 273–324.

Lichman, M. (2017). UCI Machine Learning Repository.

Luo, S. and Z. Chen (2013). Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference 143*(3), 494–504.

Luo, S. and Z. Chen (2014). Sequential lasso cum ebic for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association 109*(507), 1229–1240.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Pati, Y. C., R. Rezaiifar, and P. S. Krishnaprasad (1993). Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pp. 40–44. IEEE.

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

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 221–242.

Su, W. J. (2018). When is the first spurious variable selected by sequential regression procedures? *Biometrika 105*(3), 517–527.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association 111*(514), 600–620.

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

van de Geer, S. and J. Lederer (2013). The Bernstein–Orlicz norm and deviation inequalities. *Probability Theory and Related Fields 157*(1-2), 225–250.

van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer Science & Business Media.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association 104*(488), 1512–1524.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika 92*(4), 937–950.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics 35*(6), 2450–2473.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics 21*(1), 299–313.

Zhang, Y., J. C. Duchi, and M. J. Wainwright (2015). Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research 16*, 3299–3340.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*(Nov), 2541–2563.

APPENDIX

The online supplementary materials contain the proofs of our theoretical results, additional simulation results, and R code to reproduce our examples.

Received 9 July 2009

Accepted 8 July 2010