Carnegie Mellon University

Model Selection and Stopping Rules for High-Dimensional Forward Selection

A Dissertation Submitted to the Graduate School
in Partial Fulfillment of the Requirements for the degree

Doctor of Philosophy

in

Statistics

by

Jerzy Adam Wieczorek

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University

May 2018

For Hilary, Lukas, and Sebastian. Bank, pow, word, and scene!

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1553884. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Just as it takes a village to raise a child, I've learned that it takes a village to graduate a PhD candidate. I am deeply grateful to everyone who has been part of this journey.

I have been highly fortunate to have Jing Lei as my advisor. Jing's high standards, combined with his patient support as I strive to reach them, have tremendously influenced my growth as a researcher, writer, and academic. I look forward to our future collaborations, continuing to learn from Jing's insights on statistics (and on being a fellow parent in academia).

I am also indebted to my committee, my letter-writers, and the other CMU Statistics faculty. Larry Wasserman, Siva Balakrishnan, and Vince Vu have always been open to discussing my work and helping me see its bigger context. Ryan Tibshirani has generously invited me onto projects that allowed me to stretch in new directions. Rebecca Nugent, Chris Genovese, Joel Greenhouse, and Gordon Weinberg have enthusiastically supported my interests in teaching and pedagogy research. It was a delight to collaborate with Brian Junker and Marsha Lovett on Gen Ed assessment (and, naturally, to assess our committee as the best possible committee). When I stumbled during earlier research projects, Rob Kass, Bill Eddy, Jordan Rodu, and Anjali Mazumder mentored me through with equanimity.

I am particularly thankful to the department staff who came through on so many of my last-minute requests and hare-brained questions. Rose, Margie, Mari Alice, Laura, Carl, Heidi, Kira, Paige, Carloz, Sam, CPM, and Jess—I couldn't have kept it together without you all.

I also want to send a great big thank-you to my fellow PhD students, both in my cohort—Alex, Federico, Jisu, Justin, Lingxue, Maria, Nick, Peter, Philipp, Robert—and beyond—Brendan, Shannon, Amanda, Bret, Lee, Francesca, Jackie, Daren, Giuseppe, Spencer, Zach, Beau, Mike, Sam, Sam, and everyone else who has helped me grin and groan through our time in grad school. Nic deserves a special shout-out for stepping up as head TA and keeping my class running smoothly during these final dissertation-writing weeks.

Finally, thank you to everyone in the past who encouraged me to aim for a PhD. To my former teachers, professors, and supervisors—Katherine Bolluyt-Meints, Lynn Andrea Stein, Sarah Spence Adams, Burt Tilley, Mara Tableman, Jong Sung Kim, Robert Bertini, Jerry Maples, and Tommy Wright—thank you for your confidence in me. To my friends Mark, Andrew, and Adam—thank you for our decades (plural already!) of nerdy camaraderie.

To my parents Anna and Darek, my sister Natalia, and my entire family—thank you for a lifetime of unconditional support. Dziękuję Wam za miłość, wsparcie, i znakomite wzory do naśladowania, zarówno w pracy jak i w domu. Codziennie staram się o to, aby dotrzymać ideałów oraz przesłań naszych ukochanych dziadków, naszej Rodziny. Kocham Was.

And to my wife and sons—Hilary, Lukas, and Sebastian—your smiles, laughter, and coos make it all worthwhile. I could not, would not be here without you and your love. So so so so many I Love Yous!

Abstract

Forward Selection (FS) is a popular variable selection method for linear regression. Working in a sparse high-dimensional setting, we derive sufficient conditions for FS to attain model-selection consistency, assuming the true model size is known. Compared with earlier results for the closely-related Orthogonal Matching Pursuit (OMP), our conditions are similar but obtained using a different argument. We also demonstrate why a submodularity-based argument is not fruitful for the purpose of correct model recovery.

Since the true model size is rarely known in practice, we also 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 when using “wrapper” forward search for linear regression. This appears to be the first consistency result for any wrapper model-selection method. We illustrate intuition and demonstrate performance of our methods using simulation studies and real datasets.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Regression variable selection procedures are used widely each day to estimate sparser, more interpretable models in every quantitative field. When analyzing large, high-dimensional datasets, greedy forward selection algorithms are valued for their low computational costs and their ability to deal with the case of $p > n$. However, greedy algorithms can be difficult to study analytically, and questions remain about their model-selection consistency and practical choices of stopping rules.

Two of the most commonly used such procedures are Forward Selection (FS, Efroymson, 1960) and Orthogonal Matching Pursuit (OMP, Pati et al., 1993). To select the next variable to enter, FS finds the one additional predictor that will minimize the residual sum of squares (RSS). OMP approximates this process by merely finding the predictor most correlated with the current response residuals, as if predictors were orthogonal. For this reason, OMP has been simpler to explore analytically, while the properties of FS are less thoroughly understood. Despite conceptual similarities between FS and OMP, these procedures can differ in practice, and therefore FS deserves to be studied in its own right.

In this thesis, we study the model selection property of FS from several different perspectives. Assume iid data $(\mathbf{X}_i, Y_i)_{i=1}^n$ satisfying

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

where $\mathbf{X}_i \in \mathbb{R}^p$, and ϵ_i is independent noise with mean 0 and variance σ^2. Let $J_* = \{1 \leq j \leq p : \beta_j \neq 0\}$. First, we derive sufficient conditions under which FS will select the correct subset S of active variables after $k = |J_*|$ steps. Roughly speaking, assume that the coordinates of $\mathbf{X}$ have unit variance and absolute pair-wise correlations no larger than $1/(2k-1)$, and that the minimum absolute value of non-zero entries of β is not too small. Then we prove in Section 3.2 that

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

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). We also show that a sufficient condition for exact recovery can be derived using the property of submodularity. However, this condition is so restrictive as to be of limited practical use.

Our second main contribution is a highly practical, data-driven stopping rule for FS, using a sequential cross-validation (SeqCV) method. Consider splitting the dataset at random into two parts: a training or construction set of size n_c, and a test or validation set of size n_v, with $n_c + n_v = n$. In traditional "full" 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 as $\hat{k}$ the value of t whose $\hat{J}_t$ minimizes RSS on the test set. By contrast, in SeqCV we choose the smallest t whose $\hat{J}_t$ is a local minimizer of test RSS.

SeqCV has two advantages over FullCV at large sample sizes. First, by alternating the training and test steps, this sequential search for $\hat{k}$ can be much more efficient than FullCV when $k \ll \min\{n_c, p\}$ and the full path need not be computed. Second, SeqCV avoids FullCV's tendency to overfit. If we assume the conditions for the known-k case above, and also assume that n_c, n_v both grow quickly enough while the training ratio n_c/n_v goes to 0 quickly enough, then we show in Section 4.1 that

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

where subset $\hat{J}_{\widehat{k}}$ is selected by FS on the training data with $\hat{k}$ chosen by SeqCV. We also discuss the challenging task of selecting n_c/n_v for a given finite dataset.

Finally, in order to help analysts express the uncertainty in the model-selection process, we explore several approaches to cross-validation with confidence. By adapting general-purpose procedures for ranking with confidence, we can build confidence sets for the best model from a CV study. This also allows us a new perspective on the properties of CV with the "one standard error rule" (1SE). Simulations suggest that our proposed confidence sets have desirable properties in terms of coverage and set size, deserving further study.

1.1 Motivation

The FS algorithm is commonly used and simple to explain to non-experts. This algorithm has been a mainstay of regression textbooks since at least Draper and Smith (1966) through today (James et al., 2013; Cannon et al., 2019), even among statisticians concerned about the misuse of naive inference with FS—though Buja and Brown (2014) point a way towards valid statistical inference for FS under standard assumptions of normality. However, its properties have not been studied as completely as those of OMP and Lasso (Tibshirani, 1996), and hence this gap in the literature is worth filling for such a popular method. 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. Our theorems and large-scale simulations are a valuable update to the historical literature on FS, which consists largely of small simulation studies at low-to-moderate n and p (Dempster et al., 1977; Roecker, 1991; Derksen and Keselman, 1992; Wiegand, 2010).

Likewise, CV is a popular stopping rule among data analysts and easily explained to a lay audience, yet its properties are not thoroughly understood for random path algorithms like FS. Also, although CV is appealing because it does not make explicit distributional assumptions, this makes it difficult to know when CV is inappropriate. Finally, although CV is used to choose other algorithms' tuning parameters, CV actually has its own tuning parameters which in practice are chosen heuristically or by tradition. (These include the ratio of training to testing data; the number of folds or splits; whether to average or vote across splits; and the stopping rule variant e.g. Full CV, Sequential CV, 1SE rule, etc.) Our theorems and simulations bridge the wide gap between prior results for low-dimensional fixed-path settings vs. the modern high-dimensional random FS path setting, while illuminating the role of CV's own tuning parameters.

Further, although the property of exact recovery or correct model selection is not the only important lens through which to view FS with CV, it is another substantial literature gap. Having a true sparse linear model is often a strong assumption, rarely achieved "in the wild," and theoretical conditions for exact recovery tend to be restrictive. However, there are fields such as signal processing where both the sparsity assumption and support-recovery conditions can make sense and optimal model selection may be a reasonable goal. Outside those fields, our sufficient conditions for model selection may still help practitioners to design bigger and better future studies so that "correct" variable selection can be more plausibly approximated. Although we do not derive necessary conditions, our simulations may also help data analysts avoid using FS and CV on unsuitable already-observed datasets. Our results can indicate whether to be optimistic or pessimistic about model selection on a given dataset.

Finally, filling similar literature gaps for OMP and Lasso led to useful insights into those methods, as well as beneficial side outcomes. We have had the same experience through our work on FS. We report several unexpected but welcome contributions to our understanding of CV and related model-selection algorithms:

- Our theoretical conditions on the cross-validation training:testing split ratio inspired us to derive a large-n rule of thumb for choosing this ratio in Section 5.2. This process also allowed us to clarify why this is such a difficult problem and why the split ratio is rarely fine-tuned in practice.

- Initially, we started to study SeqCV instead of FullCV for analytical tractability. However, the large-scale simulations in Section 6.1 also turned out to demonstrate SeqCV's promising statistical performance, not to mention substantially faster computing time, as compared to FullCV.

- The benefits of combining a low training:testing ratio with SeqCV—greatly reduced computational expense with a good chance of better sparsity at comparable holdout error levels—are illustrated on real data in the Million Song Dataset example of Section 6.2.2.

- In seeking to weaken our strong sufficient condition on p for FS with SeqCV (where we conjecture it is not necessary), we proved that this condition is actually sharp for "Wrapper Forward Search" in Section 4.2. This appears to be the first consistency result for any wrapper algorithm and serves as a warning about the method's limitations in high-dimensional settings.

1.2 Related work

Forward Selection Barron et al. (2008) have studied FS under *risk consistency*: whether the risk of our estimator (selecting a model and then fitting it) converges to the oracle risk (of just fitting the best model). Also, An et al. (2008); Wang (2009) have studied FS in terms of *screening consistency*: convergence to probability 1 of choosing at least all the true predictors, but possibly also some spurious ones. Risk-consistent procedures tend to pick too-large models, while selection-consistent procedures sometimes pick too-small models (which inflates the risk), so that neither kind of consistency implies the other. To our knowledge, we provide the first conditions for *model-selection consistency* (convergence to probability 1 of exactly recovering the true predictor set) under standard FS in the high-dimensional setting.

Orthogonal Matching Pursuit Tropp (2004); Zhang (2009); Davenport and Wakin (2010); Cai and Wang (2011), and others have studied various conditions for model-selection consistency of OMP. However, OMP is not exactly the same as FS, and we illustrate the difference between them in Section 1.4.1. Also, the arguments used for OMP selection consistency are not applicable to FS due to the additional orthogonalization step. We have found that FS requires a different argument to reach the strongest possible conclusions.

Other variable selection methods Although we do not study these in detail, some connections are worth mentioning:

Both FS and OMP are greedy approximations to All-Subsets or Best Subset Regression. Foster and George (1994) show that Best Subset selection has optimal risk inflation, compared to the risk of least squares on the oracle subset if it were known. Although computing all possible subsets is combinatorially difficult for large p, recent advances in Mixed Integer Optimization have made Best Subset selection practical for much larger problems than before (Bertsimas et al., 2016). Hastie et al. (2017) run a simulation study comparing the behavior of Best Subset selection to FS and variants of the Lasso, although their focus is on prediction accuracy rather than variable selection.

Similar methods include Backward Elimination (consistency has been studied e.g. by An and Gu (1985), but it cannot be used when $p > n$) and back-and-forth Stepwise variants. An et al. (2008) show selection consistency of a complete forward path followed by a complete backward path; and Zhang (2011) shows selection consistency of a "FoBa" procedure which allows multiple backward steps after each forward addition.

Even simpler are Marginal Regression, choosing the predictors with highest marginal correlation with the response, e.g. Genovese et al. (2012); and the backwards algorithm of Zheng and Loh (1995), choosing the "most significant" predictors from a full model fit (which, again, cannot be used with $p > n$).

$L1$-regularized variable selection methods such as Lasso are motivated quite differently from FS and OMP, although Efron et al. (2004) draw a chain of connections between Lasso, Least Angle Regression (LARS), Forward Stagewise, and FS. In essence, FS takes several "large" steps: it chooses one variable at each step to maximize a correlation, then adds this variable and refits the whole model by least squares. Forward Stagewise takes many "tiny" steps, moving the selected variable only a small fraction of the way towards its least squares estimate. LARS takes intermediate steps: by going exactly the distance that Stagewise could go before another variable has larger correlation with the residuals, it is a "less greedy" variant of FS and incorporates shrinkage into our coefficient estimates. Finally, if we modify LARS to remove variables whose path passes through 0, this modified algorithm is one way to implement the Lasso. Donoho and Tsaig (2008) also show a similar chain of links between Lasso, LARS, Homotopy, and OMP. Although Lasso and LARS have theoretical and practical benefits over FS and OMP—and advances in convex optimization have made it possible to compute the Lasso path much more quickly than the FS path in many settings—it still appears that FS is more common in practice.

Meinshausen and Bühlmann (2010) introduce a concept of stability selection based on repeated subsampling and show its selection consistency in combination with the Lasso, though the method can also be applied to most of the greedy algorithms above including FS and OMP. However, they advise running around 100 subsamples, which can add considerably more computational burden than the 5 or 10 folds typical of CV.

In addition to the (continuous) variable selection problem, there is a related problem of partitioning categorical predictors with many levels. See for example the DMR algorithm of Maj-Kańska et al. (2015).

Stopping rules Practical variable-selection methods generally also require a stopping rule, which determines the final model size. Cai and Wang (2011) provide a stopping rule for OMP based on thresholding the criterion optimized in (1.1) below. Other stopping rules for FS, OMP, and Lasso have been proposed based on (adjusted) hypothesis tests and p-values, for instance Tibshirani et al. (2016); Fithian et al. (2015); or on information criteria such as AIC and BIC, e.g. Shao (1997); and on other frameworks such as minimizing the false discovery rate (FDR), e.g. Lin et al. (2012). However, all of these rules require strong distributional assumptions. Some are valid only for a fixed model set, not the random paths of FS, OMP, and Lasso.

Instead, we consider sample-splitting and cross-validation (CV), highly practical and popular methods which are valid without assuming a particular likelihood for the data or a pre-chosen model set. In an alternative to CV, Wasserman and Roeder (2009) study model selection consistency of FS, MR, and Lasso under a three-stage variant of data-splitting.

The most popular CV variants appear to be leave-one-out CV (LOOCV), V-fold CV, and Monte Carlo CV (MCCV). For V-fold CV, we partition the cases randomly into V equal-sized "folds" and cycle through them: use each fold to evaluate the model(s) trained on the remaining $V-1$ folds, then average the prediction errors across folds. LOOCV is equivalent to V-fold with $V = n$. MCCV is also similar to V-fold but uses several random splits, not a partition, so different splits' test sets may overlap. All three variants are commonly used for FullCV (choose the global minimizer of test error), which is prone to overfitting compared with our proposed SeqCV (choose the sparsest local minimizer of test error).

For model selection from a fixed model set, these CV variants have been studied for linear models by Burman (1989, 1990); Zhang (1993); Shao (1993), etc., who show that model-selection consistency requires $n_c/n_v \to 0$. This can be done straightforwardly for MCCV, but is impossible for LOOCV. To achieve this for V-fold, we must "invert" the algorithm to train on one fold and test on the others. More recently, Yang (2007) showed that the training ratio need not go to 0 if we are comparing nonparametric regression models which converge more slowly than the linear models considered earlier. Yang also distinguishes between cross-validation with averaging (as above), "CV-a," and cross-validation with voting, "CV-v." In both approaches, the prediction error for each model (or for each value of a tuning parameter such as $\hat{k}$) is computed separately on many different test splits after fitting models on the corresponding training splits. CV-a is the more common case, where the vector of test errors is averaged across splits, and we select the model (or tuning parameter) with lowest average estimated loss. Alternately, in CV-v, we pick one model (or tuning parameter value) separately on each data split; then we vote across splits, selecting the winner which had the most votes. See also the survey paper by Arlot and Celisse (2010).

Wrapper Forward Search We use FS to fit the model path and CV to choose the model size k. As an alternative, at each step, we could use the training data to fit every model with one additional variable, then evaluate them all on the test data. Such an algorithm uses CV not only as a stopping rule but also as the path-selection criterion. In the machine learning literature, this algorithm is known as Wrapper Forward Search, following terminology introduced by John et al. (1994) who distinguish between "wrapper" and "filter" methods for variable selection. Filter methods screen out variables at the start, for instance dropping predictors which have low correlation with the outcome, before training a model on the remaining predictors. In contrast, wrapper methods train models on candidate feature-subsets, then rely on holdout or cross-validation error estimates to decide which features to include. Although we focus on forward search, wrapper approaches can also be applied to backwards or back-and-forth searches.

To the best of our knowledge, we provide the first statistical model-selection consistency results for a wrapper method. Our Propositions 4.6 and 4.7 apply to Wrapper FS as well, not only to FS+SeqCV. Our Theorem 4.4 establishes a set of sufficient and (almost) necessary conditions for Wrapper FS model selection consistency.

1.3 Notation and definitions

Subscript i refers to a single observation, while c and v refer respectively to the construction (training) and validation (testing) sets. Subscripts j or h refer to a single predictor variable, while $*$, J_*, or J_h refer respectively to all columns in the true model $J_* \equiv \{1, \ldots, k\}$ or in the spurious model J_h. Unless otherwise specified, $J_h = J_* \cup h$ for some $h \notin J_*$. We use $\mathbf{X}$ for the full design matrix; $\mathbf{X}_i$ for row i; X_j for column j; 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. We use β for the full coefficient vector and β_j for element 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 $C_{j\ell} \equiv \frac{S_{j\ell}}{\sqrt{S_{jj}S_{\ell\ell}}}$.

We use vector notation for inner products and norms: $\langle a, b\rangle = a^T b$ and $\|a\|^2 = \langle a, a\rangle$.

Generic constants such as $c, c', c_1, c_2, \ldots$ do not necessarily have fixed values throughout the thesis nor even across lines within a proof.

1.4 Background

1.4.1 Model and algorithms

We assume iid data $(\mathbf{X}, Y) = (\mathbf{X}_i, Y_i)_{i=1}^n$ satisfying $Y_i = \mathbf{X}_i^T\beta + \epsilon_i$ where observations are denoted as $\mathbf{X}_i \in \mathbb{R}^p$ and ϵ_i is independent noise with mean 0 and variance σ^2, while predictor variables are denoted as $X_j \in \mathbb{R}^n$. Let $J_* = \{1 \leq j \leq p : \beta_j \neq 0\}$ and $k = |J_*|$.

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

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

where the second equality follows from a short calculation. We set $\hat{J}_{t+1} = \hat{J}_t \cup \hat{j}_{t,FS}$ and repeat, 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:

$$\hat{j}_{t,OMP} = \arg\max_{j \notin \hat{J}_t} \frac{\left|\langle Res(Y|\mathbf{X}_{\hat{J}_t}), X_j\rangle\right|}{\|Res(Y|\mathbf{X}_{\hat{J}_t})\| \cdot \|X_j\|} \,. \tag{1.1}$$

The two algorithms will take identical first steps but can differ at any later step. If we center and scale Y and all columns of $\mathbf{X}$ before the algorithm starts, OMP only needs to compute inner products and update the response residuals at each step. Meanwhile, FS must also update every unchosen predictor's residuals and rescale them at each step.

The algorithmic difference between FS and OMP can lead to practical differences. For instance, let the true model be $Y = 2X_1 + X_2$, with no noise. Let there be three predictors (X_1, X_2, X_3) to choose from, with correlations $\rho_{1,2} = 0.5$, $\rho_{1,3} = 0.25$, and $\rho_{2,3} = 0.9$. Both models correctly choose X_1 first. Then in the second step, FS correctly chooses X_2, while OMP incorrectly chooses X_3. One can also construct examples where OMP works correctly but FS does not. Since neither method strictly outperforms the other, it is worthwhile to study both.

1.4.2 Model selection consistency when k is known

When k, the number of nonzero coefficients, is known, sufficient conditions for OMP to select the correct subset of variables have been developed in Tropp (2004); Cai and Wang (2011). The main condition in these works is that the maximum pairwise correlation among the columns of $\mathbf{X}$ is smaller than $1/(2k-1)$.

The bound of $1/(2k-1)$ cannot be improved in general. To see this, consider the case where $p = k+1$, $\epsilon \equiv 0$ (noise-free), and $\beta = (1, ..., 1, 0)^T$, with the first k entries being 1 and the last one being 0. Let $\mathbf{X}$ be such that

$$X_j^T X_\ell = \begin{cases} 1 & \text{if } j = \ell\,, \\ -\mu & \text{if } j \neq \ell \text{ and } j, \ell \neq p\,, \\ \mu & \text{if } j \neq \ell \text{ and } j = p \text{ or } \ell = p\,. \end{cases}$$

Then if $\mu > 1/(2k-1)$, OMP will pick the last coordinate in the first step. Since FS and OMP choose the same variable in the first step, this example also works for FS, suggesting that the condition on μ for FS to choose the model correctly is at least as strong as that for OMP. In Chapter 3 we will show that the same condition is also sufficient for FS.

1.4.3 Stopping rules

FS+SeqCV 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 dataset at random into two parts: a training or construction set s_c of size n_c, and a test 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:

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

Choose the first model size which is a local minimizer of test MSE,

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

and select the model $\hat{J}_{\widehat{k}_{Seq}}$. We call this stopping rule sequential cross-validation (SeqCV), whether we use a single split as above, V-fold CV, or MCCV. If using V-fold CV or V splits of MCCV, we compute a separate $\widehat{MSE}_\ell\left(\hat{J}_t\right)$ on each split $\ell \in \{1, \ldots, V\}$, then choose the final model size based on $\widehat{MSE}\left(\hat{J}_t\right) = V^{-1} \sum_{\ell=1}^{V} \widehat{MSE}_\ell\left(\hat{J}_t\right)$. Our key result in Theorem 4.1 applies to both the single-split version and to MCCV with a fixed number of splits, although not to V-fold whose training ratio cannot shrink to 0.

This local-minimizer rule is similar in spirit to the "IC selection rules" of Hyun et al. (2018), who use e.g. AIC or BIC instead of test MSE to select a stopping point and then are able to condition on $\hat{k}$ in post-selection inference.

WrapperFS Several of our results below also apply to another model selection algorithm, Wrapper Forward Search (WrapperFS). With FS+SeqCV, variable selection happens on the training data alone, and CV is only used as a stopping rule. However, in WrapperFS, CV itself is used as both the variable selection mechanism and the stopping rule.

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

$$\hat{J}_t = \underset{j \notin \hat{J}_{t-1}}{\arg\min}\, \widehat{MSE}\left(\hat{J}_{t,j}\right) , \quad \hat{k}_{wrap} = \min\left\{1 \leq t \leq \min\{n_c, p\} : \widehat{MSE}\left(\hat{J}_t\right) \leq \widehat{MSE}\left(\hat{J}_{t+1}\right)\right\} ,$$

and select the model $\hat{J}_{\widehat{k}_{wrap}}$. Again, the definition above is for sample-splitting. To perform V-fold CV or MCCV instead, compute a separate $\widehat{MSE}_\ell\left(\hat{J}_{t,j}\right)$ on each split $\ell \in \{1, \ldots, V\}$, then use their average across splits $\widehat{MSE}\left(\hat{J}_{t,j}\right)$ to choose each $\hat{J}_t$ and $\hat{k}_{wrap}$.

Such wrapper search methods do not give effective test error estimates for model evaluation. However, they are reportedly quite commonly used for model selection in the data mining community because the “induction algorithm” (linear regression, neural network, decision tree, etc.) can be treated as a black box, without analytically deriving a stepwise variable-selection criterion tailored to the induction algorithm (Kohavi and John, 1997; Chrysostomou, 2009).

Chapter 2

A submodularity-based condition for exact recovery by FS

Inspired by Das and Kempe (2011), we initially attempted to study the model-selection properties of FS using the concept of submodularity.

Let f be a nonnegative set function. First, we say f is **separable** or **modular** with respect to a set U if, for any two disjoint sets S and L that are subsets of U, we have

$$\sum_{x \in S} [f(L \cup \{x\}) - f(L)] = f(L \cup S) - f(L)$$

It does not matter whether we add the elements of S piecemeal or all at once. The increase in f will be the same either way, no matter which (disjoint) S and L we use. However, if the equality $=$ is always $\geq$ instead, then we say f is **submodular**. If we choose elements to add by considering their marginal contributions, their joint improvement in f is never better than the sum of their marginal improvements.

For intuition in a statistical context, we quote Johnson et al. (2015). Submodularity implies that:

> for a set of variables A to be influential in context of another set of variables B, either A or B must be influential in isolation. Signal that is present in a complex interaction cannot be completely hidden when considering smaller sets of variables.

In our setting, let the coefficient of determination R^2 play the role of this set function f on the predictors included in our linear regression. We may wish to choose the best subset of k variables $J_{*(k)}$ to maximize the R^2 on our dataset, but this is often too computationally expensive in practice. However, if R^2 is submodular on our dataset, then greedy FS should give a decent approximation $\hat{J}_k$ to the combinatorial problem of finding $J_{*(k)}$.

In particular, a classic result from Nemhauser et al. (1978) proves that greedy algorithms cannot do too badly in the submodular case:

$$f(\hat{J}_k) \geq (1 - e^{-1}) f(J_{*(k)})\,.$$

We use this approach to seek sufficient conditions on (noise-free, fixed) datasets for FS to find the optimal model of size k. When these conditions are met, the gap between the optimal R^2 and the FS-selected model's R^2 is 0, so FS must have found an optimal model of that size.

Theorem 2.1. *Assume a fixed, noise-free dataset for which a k-sparse linear model holds exactly:* $Y_i = \mathbf{X}_i^T \beta$ *for* $i \in 1, \ldots, n$*, with* $\mathbf{X}_i \in \mathbb{R}^p$ *centered and standardized. Assume that* β *is k-sparse and that there is no sparser possible* β *which recovers* Y *exactly from* $\mathbf{X}$*. WLOG, let* β*'s nonzero entries be its first* k *entries. Also assume that* R^2 *is submodular on this dataset.*

Define $\lambda_{max}(k+m)$ *and* $\lambda_{min}(k+m)$ *to be the largest and smallest eigenvalues, respectively, of any* $(k+m) \times (k+m)$ *principal submatrix of* $\mathbf{X}^T\mathbf{X}$ *which contains the first* k *columns of* $\mathbf{X}$ *and any* m *other columns, for* $0 \leq m \leq (p-k)$*. Assume all* $\lambda_{min}(k+m)$ *are bounded from below by some positive constant for* $1 \leq m \leq k$ *(sparse eigenvalue condition). Let* $\kappa(\mathbf{X}^T\mathbf{X}) = \lambda_{max}(p)/\lambda_{min}(p)$ *be the condition number of* $\mathbf{X}^T\mathbf{X}$*. Let* β_{min} *be the smallest (in absolute value) of the* k *nonzero elements of* β*.*

Then FS will exactly recover the correct model if

$$\frac{k}{k-1} \cdot \frac{\beta_{min}^2}{\|\beta_{1:k}\|_2^2/k} > \frac{\lambda_{max}(k)}{\lambda_{min}(k+1)}\,.$$

In particular, if $\kappa(\mathbf{X}^T\mathbf{X})$ *is finite (which can happen only if* $n \geq p$ *and* $\text{rank}(\mathbf{X}) = p$*), then a sufficient condition for FS to exactly recover the correct model is*

$$\frac{k}{k-1} \cdot \frac{\beta_{min}^2}{\|\beta_{1:k}\|_2^2/k} > \kappa(\mathbf{X}^T\mathbf{X})\,. \tag{2.1}$$

Proof. The true model is $J_* = \{1, \ldots, k\}$. WLOG, we compare this against models $J = \{1+m, \ldots, k+m\}$, where $1 \leq m \leq k$ is the number of missing true variables.

Maximizing R^2 is exactly equivalent to minimizing the residual sum of squares: $RSS(J) = \sum_{i=1}^n (Y_i - \mathbf{X}_i^T \hat{\beta}_J)^2$. Let us frame the problem as maximizing the difference between the RSS for a particular model J and for the null model $\emptyset$. That is, we wish to maximize $f(J) = RSS(\emptyset) - RSS(J)$ over sets J of size k. (By assumption, no set smaller than k can be optimal.) Because R^2 is a submodular set function, f is also submodular. The minimum of f is $f(\emptyset) = 0$ and its maximum is

$$\begin{aligned} f(J_*) &= RSS(\emptyset) - RSS(J_*) \\ &= \|\mathbf{X}_{1:k}\beta_{1:k}\|_2^2 \\ &\leq \lambda_{max}(k) \cdot \|\beta_{1:k}\|_2^2 \end{aligned} \tag{2.2}$$

Recall that $Res(x|U)$ is the residual vector after projecting x onto the linear subspace spanned by the columns of U. We can lower-bound

$$\begin{aligned} f(J_*) - f(J) &= RSS(J) - RSS(J_*) \\ &= \|Res(\mathbf{X}_{1:m}\beta_{1:m}|\mathbf{X}_{1+m:k+m})\|_2^2 \\ &\geq \lambda_{min}(k+m) \cdot \|\beta_{1:m}\|_2^2 \end{aligned} \tag{2.3}$$

because

$$\begin{aligned} Res(\mathbf{X}_{1:m}\beta_{1:m}|\mathbf{X}_{1+m:k+m}) &= \mathbf{X}_{1:m}\beta_{1:m} - \mathbf{X}_{1+m:k+m}(\mathbf{X}_{1+m:k+m}^T\mathbf{X}_{1+m:k+m})^{-1}\mathbf{X}_{1+m:k+m}^T\mathbf{X}_{1:m}\beta_{1:m} \\ &= \mathbf{X}_{1:k+m}\begin{pmatrix} I_m \\ B \end{pmatrix}\beta_{1:m} \end{aligned}$$

where $B = -\mathbf{X}_{1+m:k+m}(\mathbf{X}_{1+m:k+m}^T\mathbf{X}_{1+m:k+m})^{-1}\mathbf{X}_{1+m:k+m}^T\mathbf{X}_{1:m}$ and $\left\| \begin{pmatrix} I_m \\ B \end{pmatrix}\beta_{1:m} \right\|_2 \geq \|\beta_{1:m}\|_2$.

Now, let J_t be the variable set chosen by FS by step t, with $J_0 = \emptyset$. By definition, at each step greedy FS chooses the variable with the "best" marginal contribution, maximizing $f(J_{t+1}) - f(J_t)$.

By submodularity of f, at least one marginal contribution at the first step $t = 0$ must be at least $(f(J_*) - f(J_0))/k$. There are still k unchosen correct variables before the first step, so if all of their marginal contributions were smaller than this, then they would not sum up to the optimality gap and f would not be submodular. Thus, the first variable chosen by greedy FS (whether or not it is a correct variable) must increase f by at least $1/k$ of the original optimality gap $f(J_*) - f(J_0)$, and so the optimality gap shrinks at least by a factor of $(1 - 1/k)$ during the first step:

$$f(J_*) - f(J_1) \leq (1 - 1/k) \cdot (f(J_*) - f(J_0)).$$

By the same argument, during the second step, the remaining optimality gap must shrink at least by a factor of $(1-1/k)$ if the first variable chosen was incorrect (so there are still k correct variables to be chosen) or a factor of $(1 - 1/(k-1))$ if the first variable was correct (so there are only $k-1$ correct variables left to be chosen).

Assume that after k steps, $k - m$ of the variables selected by FS were correct and the other m were incorrect, for $1 \leq m \leq k$. In the worst case, all of the incorrect variables were chosen first, so the optimality gap shrank by a $(1 - 1/k)$ factor $m+1$ times, and then by progressively changing factors:

$$\begin{aligned} f(J_*) - f(J_k) &\leq \left(1 - \frac{1}{k}\right)^m \cdot \prod_{\ell=0}^{k-m-1} \left(1 - \frac{1}{k-\ell}\right) (f(J_*) - f(J_0)) \\ &= \left(\frac{k-1}{k}\right)^m \cdot \frac{m}{k} \cdot f(J_*)\,. \end{aligned}$$

Therefore, by (2.2) and (2.3), it will be impossible to choose exactly m incorrect variables if

$$\lambda_{min}(k+m) \cdot \|\beta_{1:m}\|_2^2 > \left(\frac{k-1}{k}\right)^m \cdot \frac{m}{k} \cdot \lambda_{max}(k) \cdot \|\beta_{1:k}\|_2^2$$

or equivalently

$$\left(\frac{k}{k-1}\right)^m \cdot \min_{\substack{U \subset \{1,\ldots,k\} \\ |U|=m}} \frac{\|\beta_U\|_2^2/m}{\|\beta_{1:k}\|_2^2/k} > \frac{\lambda_{max}(k)}{\lambda_{min}(k+m)}\,.$$

The tightest case is at $m = 1$: if it is impossible to make one mistake, it also becomes impossible to make any other number of mistakes. Therefore, a sufficient condition for FS to recover the correct model is

$$\frac{k}{k-1} \cdot \frac{\beta_{min}^2}{\|\beta_{1:k}\|_2^2/k} > \frac{\lambda_{max}(k)}{\lambda_{min}(k+1)}$$

Since $\lambda_{max}(k)/\lambda_{min}(k+1) \leq \lambda_{max}(k+1)/\lambda_{min}(k+1) \leq \kappa(\mathbf{X}^T\mathbf{X})$, also sufficient is

$$\frac{k}{k-1} \cdot \frac{\beta_{min}^2}{\|\beta_{1:k}\|_2^2/k} > \kappa(\mathbf{X}^T\mathbf{X})$$

when $\mathbf{X}$ has full column rank. □

However, the condition above is too limited to be of much use, for two reasons:

a. Even after assuming submodularity, our sufficient condition for success of FS is very restrictive. By definition, $\kappa(\mathbf{X}^T\mathbf{X}) \geq 1$, and $\beta_{min}^2 \leq \|\beta_{1:k}\|_2^2/k$. Thus, even if $\kappa(\mathbf{X}^T\mathbf{X}) = 1$, the elements of $\beta_{1:k}$ must be all close in magnitude: $\beta_{min}^2 \in \left(\frac{k-1}{k}, 1\right) \cdot \|\beta_{1:k}\|_2^2/k$. Alternately, even if $\beta_{1:k}$ has constant entries, the condition number must be small: $\kappa(\mathbf{X}^T\mathbf{X}) \in \left(1, \frac{k}{k-1}\right)$.

b. The known conditions for R^2 to be submodular are also very restrictive. Das and Kempe (2011)'s condition for R^2 to be submodular is equivalent to requiring the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}$ to be 1, for $\mathbf{X}$ with normalized columns. In other words, this condition requires orthogonal $\mathbf{X}$, in which case the success of FS is trivial anyway.

We could ease concern (b) by using Das and Kempe (2011)'s concept of "weak" or "approximate submodularity." The authors define a concept they call the submodularity ratio, γ, such that submodularity holds iff $\gamma \geq 1$. The situation where $\gamma \in (0, 1)$ is called weak submodularity. In this setting, Das and Kempe use γ to bound the optimality gap after running FS, OMP, or Marginal Regression for k steps, finding conditions where the greedy solution is not much worse than the (combinatorially-difficult-to-find) optimal solution. Additionally, Elenberg et al. (2017) extend the work of Das and Kempe by defining a new concept of "restricted strong convexity" which implies weak submodularity. They show that by running FS for $r > k$ steps, we can achieve R^2 arbitrarily close to that of the optimal solution of size k (and, under further assumptions, the ℓ_2 norm of the error in estimating β goes to 0) if r is of order $\log(n)$ and weak submodularity holds.

However, this merely shows that if weak submodularity holds, FS can find *some* too-large model that approximates the truth well. In the context of *correct* support recovery, using weak submodularity to deal with concern (b) would make concern (a) even worse, as would extending our results to allow for non-zero noise and/or random $\mathbf{X}$. We did not pursue these directions further.

Finally, we use the sufficient condition in (2.1) to derive an incoherence condition, for comparison with our other results below. Define the coherence μ as the greatest absolute correlation between any two columns of $\mathbf{X}$. Then the condition number is bounded by $\kappa(\mathbf{X}^T\mathbf{X}) \leq \frac{1+(k-1)\mu}{1-\mu}$. Meanwhile, the loosest bound for the left-hand side of (2.1) is when the nonzero elements of β are all equal. Then a sufficient condition would be

$$\frac{k}{k-1} > \frac{1+(k-1)\mu}{1-\mu} \quad \Rightarrow \quad \frac{1}{(k-1)^2+k} > \mu\,.$$

This requires μ on the order of k^{-2} or smaller, which is much stricter than the incoherence condition of $\mu < (2k-1)^{-1}$ that we derive in Chapter 3. Again, submodularity is not as fruitful as other approaches to tackling this problem.

Chapter 3

Path consistency of FS

Throughout this chapter, we study conditions for the path-consistency of FS—equivalently, the model-selection consistency of oracle FS with known k. First, Section 3.1 summarizes our main conclusions and proof approach for the case of fixed predictors and noise. Next, Section 3.2 states our results for the random case, along with discussion of related conditions for OMP and the Lasso. Finally, Section 3.3 contains the formal statements of our fixed-data claims, along with supporting results and extensions for both the fixed- and random-data cases. Non-trivial proofs are deferred to Chapter 8.

3.1 Fixed data

For now, assume the k-sparse linear model from Section 1.4 holds with fixed design matrix and noise vector. Let every predictor column X_j be standardized to zero mean and unit variance: $\|X_j\|^2/n = 1$. Define the *coherence among predictors* as $\mu = \max_{j \neq \ell \in \{1,\ldots,p\}} n^{-1}|X_j^T X_\ell|$.

First, if the true model is noiseless so that $\epsilon \equiv 0$, then by Corollary 3.6, a sufficient condition for correct model selection at each step $t \geq 2$ is

$$\sqrt{\frac{1 - t\mu}{1 - (t-1)\mu}} > (2k - 2t - 1)\frac{\mu}{1 - (t+1)\mu}$$

which is implied by $\mu < (2k-1)^{-1}$ with $k \geq t+2$. To derive this condition, we consider the QR decomposition of the design matrix: $n^{-1/2}\mathbf{X} = ZA$, where Z has orthonormal columns and A is an upper triangular matrix with entries $a_{j\ell}$. We also note that $A^T A$ is the Cholesky decomposition of $n^{-1}\mathbf{X}^T\mathbf{X}$. Using matrix perturbation results from a theorem of Sun (1992), we derive lower and upper bounds on $|a_{j\ell}|$ in terms of the coherence μ. If we assume FS has made correct choices so far up to step t, the conditions for FS' next step to be correct can be written in terms of Z and $a_{j\ell}$, and plugging in the bounds on $|a_{j\ell}|$ gives the sufficient condition above.

Next, for a fixed but nonzero noise vector ϵ, define the *coherence between noise and predictors* as $\gamma = \max_{j\in\{1,\ldots,p\}} n^{-1/2}|X_j^T(\epsilon/\|\epsilon\|)|$. Proposition 3.3 repeats the argument above but now also accounting for γ, and Corollary 3.4 summarizes the results in a "beta min" condition: If $\mu, \gamma < (2k-1)^{-1}$, then FS will choose the correct model if

$$\min_{j\in 1\ldots,k} \frac{|\beta_j|}{\|\epsilon\|} \geq \frac{16.8k\gamma}{\sqrt{n}} .$$

Since $k\gamma$ is at most constant, this lower bound on the signal-to-noise ratio (SNR) is at most of order $n^{-1/2}$.

Finally, if the design is orthogonal, we can drop the dependence on k. By Corollary 3.5, if $\mu = 0$, then FS will choose the correct model if

$$\min_{j\in 1\ldots,k} \frac{|\beta_j|}{\|\epsilon\|} \geq \frac{2.28\gamma}{\sqrt{n}} .$$

3.2 Random data

From now on, we will assume that the data are random and come from a sub-Gaussian distribution, i.e. one whose tails decay at least as fast as Gaussian tails. 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 $\|u\|_2 = 1$, $\|u^T V\|_{\psi_2} \leq c < \infty$.

This definition uses the Orlicz ψ-norm: For a univariate random variable Z,

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

Specifically, we use $\psi_2(x) = e^{x^2} - 1$. 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 $\mathbb{V}(\mathbf{X}_i) = \Sigma$ and $\mathbb{V}(\epsilon_i) = \sigma^2$. Σ is positive definite. Without loss of generality, Σ is a correlation matrix (has 1s along the diagonal).*

Throughout Chapter 3 we can weaken Assumption 1 to allow only uncorrelated $\mathbf{X}$ and ϵ. Their independence will be required in Chapter 4. Also, if Σ is not a correlation matrix, then by rescaling $\mathbf{X}$ and β appropriately, we can still apply these results to predictors whose covariance matrix has other diagonals, as long as the diagonal entries are uniformly bounded in n by some finite positive constant.

Assumption 2. *The true model (denoted J_*) is k-sparse, and the signal is contained in the first k covariates, so that $Y = \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$ with $|\beta_1| \geq \ldots \geq |\beta_k| > 0$.*

If Σ is scaled to be a correlation matrix with entries $\rho_{j\ell}$, then let $\mu = \max_{j \neq \ell} |\rho_{j\ell}|$. We call μ the *population coherence among predictors.*

Assumption 3. *Let p and k grow*[*] *with n. As $n, p, k \to \infty$, we require that:*

1. *the population coherence must shrink as k grows: $\mu < (2k-1)^{-1}$;*
2. *the signal may shrink but not too quickly: $|\beta_{min}| \geq c \cdot k\sigma\sqrt{\frac{\log(p)}{n}}$ for some constant $c > 0$;*
3. *the dimension cannot grow too much faster than sample size: $n^{-1} \cdot \sigma^2 k^2 \log(p) \to 0$.*

Theorem 3.1. *Assume 1, 2, and 3. Then oracle FS is model-selection consistent:*

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

In particular, under these conditions and with k known, FS chooses the correct model J_* with probability at least $1 - c'p^{-\eta}$ for some $c' > 0$ and our choice of $\eta > 0$.

The proof of Theorem 3.1 is a direct consequence of Proposition 3.10 and Corollary 3.4, which are given in Section 3.3 and proven in Chapter 8. Our proof proceeds from the fixed-noise, fixed-design case 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 $\epsilon \sim N(0, \sigma^2 I_n)$. They derive model-selection consistency of OMP if $\mu < (2k-1)^{-1}$ and, for any $\eta \geq 0$,

$$\min_{i \in 1,\ldots,k} |\beta_i| \geq \frac{2\sigma\sqrt{2(1+\eta)\log p}}{1-(2k-1)\mu}\,. \tag{3.1}$$

If we assume the columns of $\mathbf{X}$ have unit variance instead of unit norm, we can replace $|\beta_i|$ with $\sqrt{n}|\beta_i|$ in (3.1) 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, very similar to Tropp's ERC and also implied by $\mu < (2k-1)^{-1}$, that is sufficient and "almost necessary" for model-selection consistency. As summarized in Section 2.6 of Bühlmann and van de Geer

[*] 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 $\log(N)$, and we can replace probabilities of the form $1 - cp^{-\eta}$ with $1 - cN^{-\eta}$.

(2011), if we have this condition and also sufficiently strong signal $|\beta_{min}| \gg \sqrt{k\log(p)/n}$, then there exists a sequence of Lasso regularization parameters λ_n for which the Lasso is model-selection consistent.

Recall from Section 1.4.2 that the incoherence condition $\mu < (2k-1)^{-1}$ is sharp among conditions based only on coherence: there exist cases where the condition is not just sufficient but necessary. Disappointingly, 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 set of true variables, while incoherence also restricts all correlations among spurious variables.

On the other hand, our simulations in Section 6.1 suggest that this condition is not usually necessary unless Σ has a particularly disadvantageous structure. Besides, incoherence can be approximately checked in practice with good estimates of μ and k. The ERC or IC cannot be checked without knowing the true support J_*, even though our goal in model selection is to learn J_*. Finally, we can relax the incoherence condition a little if we assume that Σ is row-sparse.

Corollary 3.2. *Assume that each row of Σ is s-sparse off of the diagonals, i.e. has s nonzero off-diagonal entries. Let $1 \leq s < k$. Assume 1, 2, and a modification of 3, replacing $\mu < (2k-1)^{-1}$ by $\mu < (3.4s)^{-1}$. Then oracle FS is model-selection consistent.*

The result follows directly from Proposition 3.10 with Corollary 3.7. If $s \ll k$, the new condition that $\mu < (3.4s)^{-1}$ can be much less restrictive than the original requirement that $\mu < (2k-1)^{-1}$.

3.3 Supporting results

A comment on notation: in Sections 3.3.1 and 3.3.2, we assume the columns of $\mathbf{x}$ are fixed and standardized to unit norm: $x_j \equiv \frac{X_j - \overline{X}_j}{\|X_j - \overline{X}_j\|}$. In Section 3.3.3, we will return to random $\mathbf{X}$ with columns of unit variance.

3.3.1 Fixed design and noise

Assumption 4. *Let $\mathbf{x}$ be a fixed $n \times p$ matrix, with each column normalized to zero mean and unit norm, and define $\Sigma = \mathbf{x}^T\mathbf{x}$. Let E be a fixed n-vector, not necessarily normalized, and let the true model be k-sparse. WLOG assume that the first k covariates are the nonzeros, and the coefficients are ordered: $y = \beta_1 x_1 + \ldots + \beta_k x_k + \mathrm{E}$, with $|\beta_1| \geq \ldots \geq |\beta_k| > 0$.*

Define the *coherence among predictors* as $\mu = \max_{j \neq \ell \in \{1,\ldots,p\}} |x_j^T x_\ell|$. Define the *coherence between noise and predictors* as $\gamma = \max_{j \in \{1,\ldots,p\}} |x_j^T(\mathrm{E}/\|\mathrm{E}\|)|$.

Proposition 3.3. *Assume 4 and let $\mu < (2k-1)^{-1}$. Then these are sufficient conditions for FS to select a correct term at each given step, if all previous steps have also been correct. For $t = 0$,*

$$\frac{|\beta_1|}{\|\mathrm{E}\|} > \frac{2\gamma}{1-(2k-1)\mu}$$

then for $t = 1$,

$$\frac{|\beta_2|}{\|\mathrm{E}\|} > \frac{2\gamma}{1-(2k-2)\mu}$$

and for $t = 2, \ldots, k-1$,

$$\frac{|\beta_{t+1}|}{\|\mathrm{E}\|} > \frac{\frac{2\gamma}{1-t\mu-(t+1)\gamma^2}}{\sqrt{\frac{1-t\mu}{1-(t-1)\mu}} - \frac{(2k-2t-1)\mu}{1-(t+1)\mu}}\,.$$

The following corollary gives a general "beta min" condition for all steps of FS to succeed:

Corollary 3.4. *Assume 4. If we have both $\gamma, \mu < (2k-1)^{-1}$, then FS will choose the correct model if the signal-to-noise ratio is at least*

$$\min_{i \in 1,\ldots,k} \frac{|\beta_i|}{\|\mathrm{E}\|} \geq 16.8k\gamma\,.$$

Also, an equivalent result clearly holds if we rescale the data and noise (but not the coefficients) by $\sqrt{n}$. Let $(Y, \mathbf{X}, \epsilon) = \sqrt{n}(y, \mathbf{x}, \mathrm{E})$, so that each column X_j has unit variance: $\|X_j\|^2/n = 1$, so $Y = \mathbf{X}\beta + \epsilon$, with β as before. Then FS will choose the correct model if

$$\min_{j \in 1,\ldots,k} |\beta_j| \geq 16.8k\gamma\|\epsilon\|/\sqrt{n}\,.$$

We follow up with several extensions to special cases.

First, if the design is orthogonal, we can drop the dependence on k:

Corollary 3.5. *Assume 4. If $\mu = 0$, FS will choose the correct model if the signal-to-noise ratio is at least*

$$\min_{j \in 1,\ldots,k} \frac{|\beta_j|}{\|\mathrm{E}\|} \geq \frac{25}{11}\gamma \approx 2.28\gamma\,.$$

Proof. Directly from Proposition 3.3, we have the sufficient condition

$$\min_{j \in 1,\ldots,k} \frac{|\beta_j|}{\|\mathrm{E}\|} \geq \frac{2\gamma}{1-k\gamma^2}\,.$$

The result follows from assuming $\gamma < (2k-1)^{-1}$ and maximizing the denominator above at the "worst case" of $k = 3$. (If $k \leq 2$, then 2γ is a sufficient lower bound.) □

Next, if there is no noise term, we can drop the $|\beta_{min}|$ lower-bound altogether:

Corollary 3.6. *Assume 4. If the true model is noiseless, a sufficient condition at each step $t \geq 2$ is*

$$\sqrt{\frac{1-t\mu}{1-(t-1)\mu}} > (2k-2t-1)\frac{\mu}{1-(t+1)\mu}$$

which is implied by $\mu < (2k-1)^{-1}$ with $k \geq t+2$.

Proof. Follow the same argument as in the proofs of Proposition 3.3 and Corollary 3.4, i.e. Cholesky decomposition and correlation matrix inversion, but with noise vector $\mathrm{E} \equiv 0$. □

Finally, if we assume that Σ is not orthogonal but row-sparse, we may be able to allow larger coherence values μ:

Corollary 3.7. *Assume 4. Additionally, assume that each row of Σ is s-sparse off of the diagonals, i.e. has s nonzero off-diagonal entries. Let $1 \leq s < k$.*

If $\mu < (3.4s)^{-1}$ and $\gamma < \sqrt{\frac{12}{17k}}$, then FS chooses the correct model if the signal-to-noise ratio is at least

$$\min_{j \in 1,\ldots,k} \frac{|\beta_j|}{\|\mathrm{E}\|} \geq \frac{\gamma \cdot q(s)}{\frac{12}{17} - k\gamma^2}$$

where $q(s) \equiv 2 \cdot \left(\sqrt{\frac{2.4s}{2.4s+1}} - \frac{2}{2.4}\right)^{-1}$ is greatest at $q(1) \approx 293$ but asymptotes towards $q(s) \approx 12$ as $s \to \infty$.

The coherence condition of Corollary 3.7 is less restrictive than that of Corollary 3.4 only if $s \ll k$.

3.3.2 Fixed design, random noise

Assumption 5. *Define $\mathbf{x}$, $\{\beta_1, \ldots, \beta_k\}$, and $y = \beta_1 x_1 + \ldots + \beta_k x_k + \mathrm{E}$ as in Assumption 4, but now assume that each element of E has variance σ^2/n and is i.i.d. from some sub-Gaussian distribution.*

Let $\hat{\gamma}$ denote the observed coherence between the sample noise and predictors.

Proposition 3.8. *Assume 5. For any choice of $\eta > 0$, $\hat{\gamma}\|\mathrm{E}\| \equiv \max_{j \in 1,\ldots,p} |x_j^T \mathrm{E}| = O(\sigma\sqrt{\log(p)/n})$ with high probability (at least $1 - c' p^{-\eta}$ for some $c' > 0$).*

Corollary 3.9. *Assume 5. If $\mu < (2k-1)^{-1}$ and $\sigma^2 k^2 \log(p)/n \to 0$, then $\exists\ c > 0$ s.t. FS chooses the correct model with high probability (at least $1 - c' p^{-\eta}$ for some $c' > 0$) if the signal-to-noise ratio is at least*

$$\min_{j \in 1,\ldots,k} \frac{|\beta_j|}{\sigma} \geq ck\sqrt{\log(p)/n}$$

where c depends on our choice of η, and both c, c' depend on the particular sub-Gaussian distribution of E. Also, the same result clearly holds if we rescale the data and noise (but not the coefficients) by $\sqrt{n}$. Let $(Y, \mathbf{X}, \epsilon) = \sqrt{n}(y, \mathbf{x}, \mathrm{E})$, so that each element of ϵ has variance σ^2, and each column X_j has unit variance: $\|X_j\|^2/n = 1$. Then $\hat{\gamma}\|\epsilon\| \equiv \max_{j \in 1,\ldots,p} |(X_j/\|X_j\|)^T \epsilon| = \sqrt{n}\hat{\gamma}\|\mathrm{E}\|$, so that $\hat{\gamma}\|\epsilon\|/\sqrt{n} = O(\sigma\sqrt{\log(p)/n})$.

Proof. The result follows directly from Proposition 3.8 and Corollary 3.4. The condition that $\sigma^2 k^2 \log(p)/n \to 0$ ensures that $\hat{\gamma} < (2k-1)^{-1}$ with high probability for Corollary 3.4. □

3.3.3 Random design and noise

Now use the assumptions and setup of Section 3.2, where the columns of $\mathbf{X}$ have unit variance. Let $\hat{\mu}$ denote the observed sample coherence among the predictors.

Proposition 3.10. *Assume 1 and 2. Let $\mu < (2k-1)^{-1}$ and $\sigma^2 k^2 \log(p)/n \to 0$.*

For any choice of $\eta > 0$ and sufficiently large n, with high probability (at least $1 - cp^{-\eta}$ for some $c > 0$) we have jointly that $\hat{\gamma}\|\epsilon\|/\sqrt{n} \equiv \max_{j \in 1,\ldots,p} \left| \frac{(X_j - \overline{X}_j)^T \epsilon}{\|X_j - \overline{X}_j\|} \right| = O(\sigma\sqrt{\log(p)/n})$ and that both $\hat{\gamma}, \hat{\mu} < (2k-1)^{-1}$.

Chapter 4

Model-selection consistency of FS+SeqCV

In Chapter 3 we assumed that the correct model size k is known. Here, we assume more realistically that k must be estimated instead. Consider estimating k by FS with sequential data-splitting, or FS+SeqCV, as defined in Section 1.4. In Section 4.1 we show that this procedure is also model-selection consistent, under the additional conditions below. Next, Section 4.2 explores one of these conditions—a restrictive requirement that the dimension grows less quickly than the square root of the training sample size. We conjecture that this condition is not necessary for FS+SeqCV, but we prove that it is sharp for the closely-related Wrapper FS algorithm. Finally, Section 4.3 contains the formal statements of our supporting results, whose proofs are deferred to Chapter 8.

4.1 Sufficient conditions

Assumption 6. *As $n, p, k \to \infty$, we require that:*

1. *the conditions of Assumption 3 hold on the training sample, with n replaced by n_c;*
2. *the coefficients must not be too unbalanced: $\frac{\beta^2_{min}}{\beta^2_{max}} \geq c \cdot \max\left\{k\sqrt{\frac{\log(k)}{n_c}}, \frac{k^2\log(k)/n_v}{\beta^2_{min}/\sigma^2}\right\}$ for some constant $c > 0$;*
3. *the dimension cannot grow as quickly as the training sample size or the test-train ratio: $\min\left\{\frac{k}{n_c}, \frac{n_c}{n_v}\right\} \cdot kp^2\log(p) \to 0$.*

The balanced-coefficients condition 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. 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.

Proposition 4.6 shows that $B_h \approx \tilde{\beta}^2_{J_h} + O_p\left(n_c^{-3/2}\right)$, where $\tilde{\beta}_{J_h}$ is the regression coefficient for the noise regressed on X_h after projecting out $\mathbf{X}_*$. Our required rate of $p^2/n_c \to 0$ comes from a careful analysis of the expansion of B_h which exploits a cancellation between the $\tilde{\beta}^2_{J_h}$ and the $O_p(n^{-3/2})$ terms. In contrast, a straightforward argument directly using rates of convergence for $\tilde{\beta}^2_{J_h}$ 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. By requiring $\frac{n_c}{n_v} \to 0$, we can prevent overfitting by ensuring that the worst-case difference in risks is larger than the test-set error in estimating this difference. Roughly speaking, Proposition 4.7 shows that the additional condition $|\tilde{\beta}_{J_h}| > n_v^{-1/2}$ prevents a model-selection mistake if we have already avoided a training mistake:

$$\mathbb{P}(\text{selection mistake for } h) \lesssim \mathbb{P}\left(|\tilde{\beta}_{J_h}| < n_v^{-1/2}\right) = \mathbb{P}\left(\sqrt{n_c}|\tilde{\beta}_{J_h}| < \sqrt{n_c/n_v}\right) \asymp \sqrt{n_c/n_v}\,.$$

By a union bound argument, $p\sqrt{n_c/n_v} = o(1)$ is sufficient.

Note that Assumptions 1, 2, 6 together satisfy the conditions of Theorem 3.1 on the training sample, so that FS is path-consistent under these conditions.

Theorem 4.1. *Assume 1, 2, 6. Then sample-splitting with FS+SeqCV is model-selection consistent:*

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

Proof. By Proposition 4.5, with probability approaching 1, FS+SeqCV will select a correct model path and will not stop before finding model J_*.

Then, the next comparison will be between the true model and one of the $p - k$ spurious models $Y = \beta_1 X_1 + \ldots + \beta_k X_k + \beta_h X_h + \epsilon$, for candidate covariate $h \in k+1, \ldots, p$. By Proposition 4.7, with probability approaching 1, model J_* will have lower test MSE than any of these $p - k$ spurious models, so FS+SeqCV must stop at model J_*. □

For comparing nested underfitting models J_h and $J_{h'}$ on a correct model path, Proposition 4.5 decomposes the difference $\widehat{MSE}(J_h) - \widehat{MSE}(J_{h'})$ into signal and noise components, then derives conditions for the signal to be detectable (so that the smaller model is not chosen and FS+SeqCV does not stop before the true model).

For comparing the true model J_* with any one-term-larger spurious model J_h, Proposition 4.6 makes B_h explicit and derives conditions under which the probability of a training mistake vanishes. Proposition 4.7 extends this argument to cover the probability of an overall model-selection mistake.

Our arguments rely on a union bound to protect against the worst case over all $p - k$ possible spurious models, not just the single spurious model that FS chooses to train and test. When using FS to determine the model path, we conjecture in Section 4.2 that the last line of Assumption 6 could be weakened. Even so, Proposition 4.7 is of independent interest, since its worst-case setup corresponds to the Wrapper FS algorithm defined in Section 1.4. In Section 4.2, we show that Assumption 6.3 is not only sufficient but also "almost" necessary for Wrapper FS—the condition is sharp in the sense that there are situations where the Assumption cannot be weakened.

So far we have assumed a single data-split, but it is more common to perform cross-validation by combining estimates of loss by averaging (CV-a) or voting (CV-v) across many repeated splits of the same dataset. Our results extend straightforwardly to Monte Carlo CV (MCCV) with a shrinking test-train ratio.

Corollary 4.2. *Under the conditions of Theorem 4.1, FS+SeqCV-v is also model-selection consistent. That is, instead of running FS+SeqCV on a single split, we can run it on all possible splits with the same ratio n_c/n_v (or any random subset of MCCV splits), then vote across the estimated model-selections, and choose the single model with the most votes.*

Corollary 4.2 follows from our Theorem 4.1 and the argument in Theorem 2 of Yang (2007). Note that we still require $n_c/n_v \to 0$; Yang was able to allow the training set to dominate only when comparing nonparametric models, which converge at a slower rate than our linear models.

Corollary 4.3. *Under the conditions of Theorem 4.1, FS+SeqCV-a is also model-selection consistent for any fixed number of MCCV splits. That is, instead of running FS+SeqCV on a single split, we can run it on a random subset of all possible splits with the same ratio n_c/n_v. Record $\widehat{MSE}$ for each model across the splits, and choose the single model with the lowest average $\widehat{MSE}$ across splits.*

If FS+SeqCV tends to choose the right model with probability going to 1, it will do so on each of the CV-a splits. Hence, with high probability, the true model will have lowest MSE on every split, and therefore lowest average MSE across splits. A union bound takes care of the fact that splits on the same dataset are not independent.

Yang (2007) gives intuition for why CV-v ought to perform similarly to CV-a for sufficiently high signal-to-noise ratio, but the differences between them appear to be second-order effects that are difficult to analyze theoretically. However, both CV-a and CV-v intuitively should (and empirically appear to) perform better at any finite n than single-split CV does, and CV-a appears to outperform CV-v in moderate-n simulations.

Our own simulations show similar performance no matter whether CV-v or CV-a is used, and no matter whether Monte Carlo CV or V-fold CV is used. For brevity, Chapter 6 reports simulations and examples only for the most commonly taught and used variant: V-fold CV-a (as well as inverted V-fold, to let the test-train ratio shrink).

4.2 On the effect of p

Although we conjecture that $p^2 \ll n_c$ is not strictly necessary for FS+SeqCV, it is still a meaningful condition for CV in high-dimensional regression. This condition appears in Theorem 4.1 through Propositions 4.6 and 4.7, which use a worst-case union-bound as if the algorithm were evaluating all $p-k$ spurious models. This worst-case setup is stricter than FS requires, but it corresponds exactly to the Wrapper FS algorithm defined in Section 1.4.

To the best of our knowledge, we provide the first statistical model-selection consistency results for a wrapper method. Previous work as summarized in Chrysostomou (2009) has only evaluated performance using simulations or focused on computational speed-ups. Furthermore, we can show that the condition $p^2 \ll n_c$ is not only sufficient for Wrapper FS but also "almost" necessary, in that breaking this condition prevents model-selection consistency even under a very simple setup.

Consider the case of independent Gaussian data and noise, orthogonal design matrix, and constant-mean true model. Theorem 4.4 shows that if $p^2/n_c \not\to 0$, then even in this simple case there is a non-vanishing probability that the trained true model's risk can be beaten by *some* spurious trained model, which is carried over to the test-set estimates of risk, resulting in overfitting. So the condition $\frac{p^2}{n_c} \to 0$ in Assumption 6.3 is necessary for selection consistency of Wrapper FS in this setting.

Assumption 7. *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 $h \in 1, \dots, p$.*

Assumption 8. *$\mathbf{X}_{n\times p}$ 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 9. *As $n \to \infty$, the number of candidate variables p grows "too quickly": $\liminf p^2/n_c \geq \Gamma$ for some constant $\Gamma > 0$. It does not matter whether or not the training/testing split ratio n_c/n_v goes to 0.*

Theorem 4.4. *Assume 7, 8, 9. Then the probability that Wrapper FS makes a model-selection mistake does not vanish:*

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

Theorem 4.4 is proved in Chapter 8. First we prove Proposition 4.8, which shows the non-vanishing probability of a training mistake. Next, the proof of Theorem 4.4 shows that no choice of testing sample size can make the probability of a mistake vanish.

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 Wrapper FS, but FS+SeqCV will rarely require that same bound in practice. FS will only test the single spurious model with the lowest *training-data estimate* of risk, which will have one of the highest *true* risks if $p - k$ is large; so FS+SeqCV should not overfit after finding the true model. Figure 4.1 illustrates this intuition: the top row shows a plausible outcome with a single spurious variable, while the bottom row shows a typical outcome when FS chooses the best training-data fit from among many spurious variables.

We also illustrate these arguments with simulations to compare FS+SeqCV and Wrapper FS. We draw n sample outcomes from a true null model (fixed intercept and constant-variance Normal errors), along with p spurious predictors from an orthogonal random Normal design (zero mean and identity covariance matrix). For the Wrapper FS and FS+SeqCV methods, respectively, the top and bottom subplots of Figure 4.2 show heatmaps of $\mathbb{P}$(correctly choose null model) at different combinations of p and n_c, estimated from 500 replications at each combination. In order to evaluate both conditions of interest using a single pair of figures, we choose $n_c = \sqrt{n}$ so that $\frac{n_v}{n_c} = \sqrt{n} - 1 \approx n_c$.

In the top of Figure 4.2, the estimated contours of constant success probability are roughly shaped like $p = \sqrt{n_c} \approx \sqrt{\frac{n_v}{n_c}}$ when using Wrapper FS. As we expect from Theorem 4.4, model selection is challenging at high p; becomes easier at high n; and tends towards success probability 1 only if n_c and $\frac{n_v}{n_c}$ grow much faster than p^2 does.

However, the bottom of Figure 4.2 shows a completely different pattern when using FS+SeqCV. Success probability increases with n_c at every p, but also with p at every n_c. By using FS on the training set to select the next candidate model, correct selection also gets easier as p grows at *every* n, as long as n_c and $\frac{n_v}{n_c}$ grow. Therefore, it seems reasonable to conjecture that FS+SeqCV is consistent even if the p^2 factor is removed from Assumption 6.3. We leave this as a focus for future work.

4.3 Supporting results

4.3.1 $\mathbb{P}(underfit) \to 0$

Proposition 4.5. *Assume 1, 2, 6. Then FS+SeqCV will construct a correct model path on the first k steps and will not stop before step k, with probability at least* $1 - \big(cp^{-1} + c'(k^{-1} + (2/e)^k)(1 - cp^{-1})\big) \to 1$.

4.3.2 $p^2/n_c \to 0$ is sufficient for $\mathbb{P}(overfit) \to 0$

For a given training dataset and spurious covariate h, recall from Section 4.1 that B_h is the expectation (over possible test datasets) of the difference in test MSE estimates between the true model J_* and the spurious

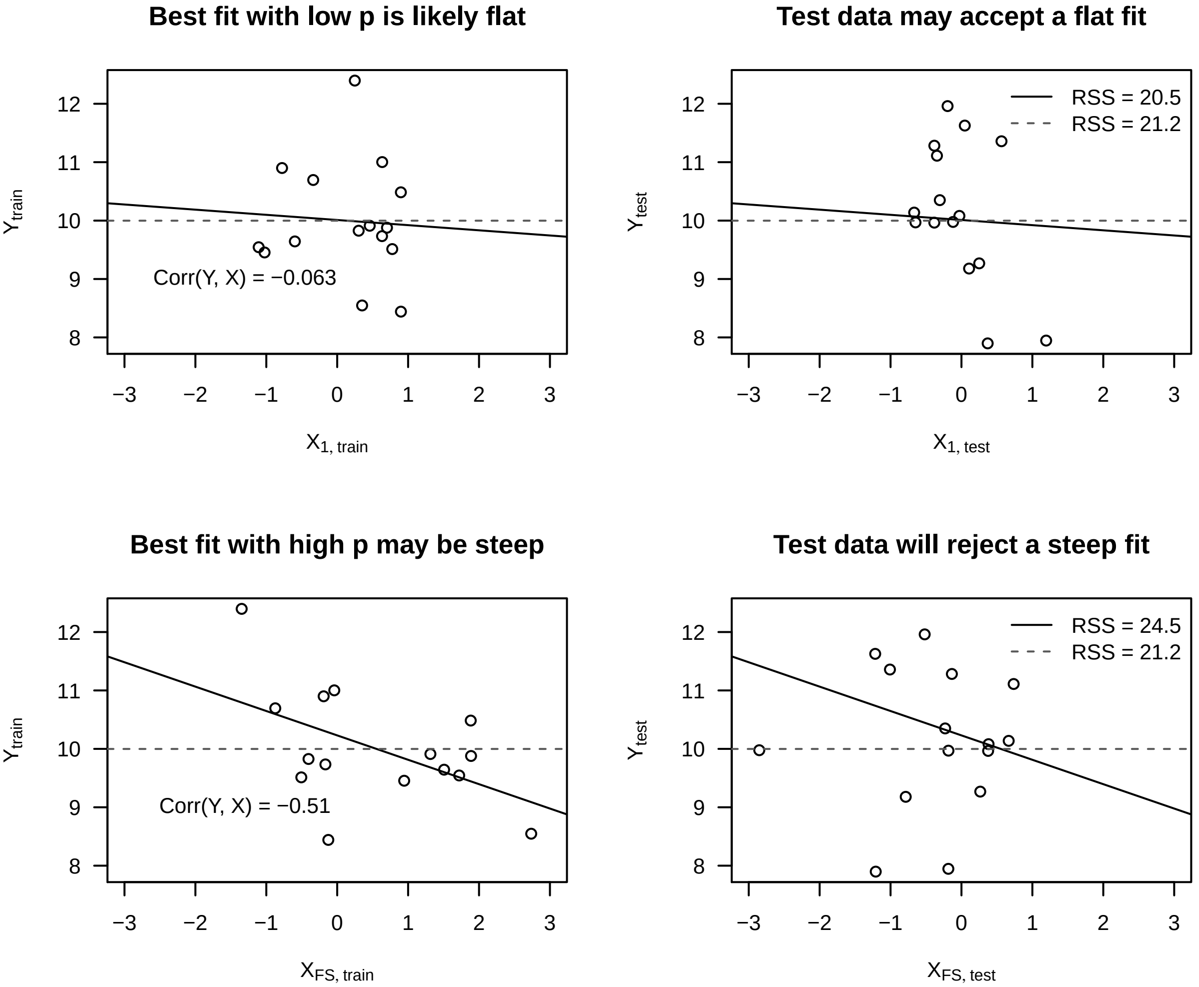

Figure 4.1: Example simulations from a true null model. Top row: if we have only a few spurious variables to choose from, the best training-data fit is probably quite flat. By bad luck, it might fit well to the test data too, leading us to select this overfitting model. Bottom row: if there are many spurious variables to choose from, the best training-data fit will be quite steep, overfitting to the training data. This will most likely be rejected by the test data in favor of the null model's flat fit, preventing FS+SeqCV from overfitting at higher p. However, WrapperFS will choose the spurious variable whose trained slope fits the test data best, so overfitting becomes more likely at higher p.

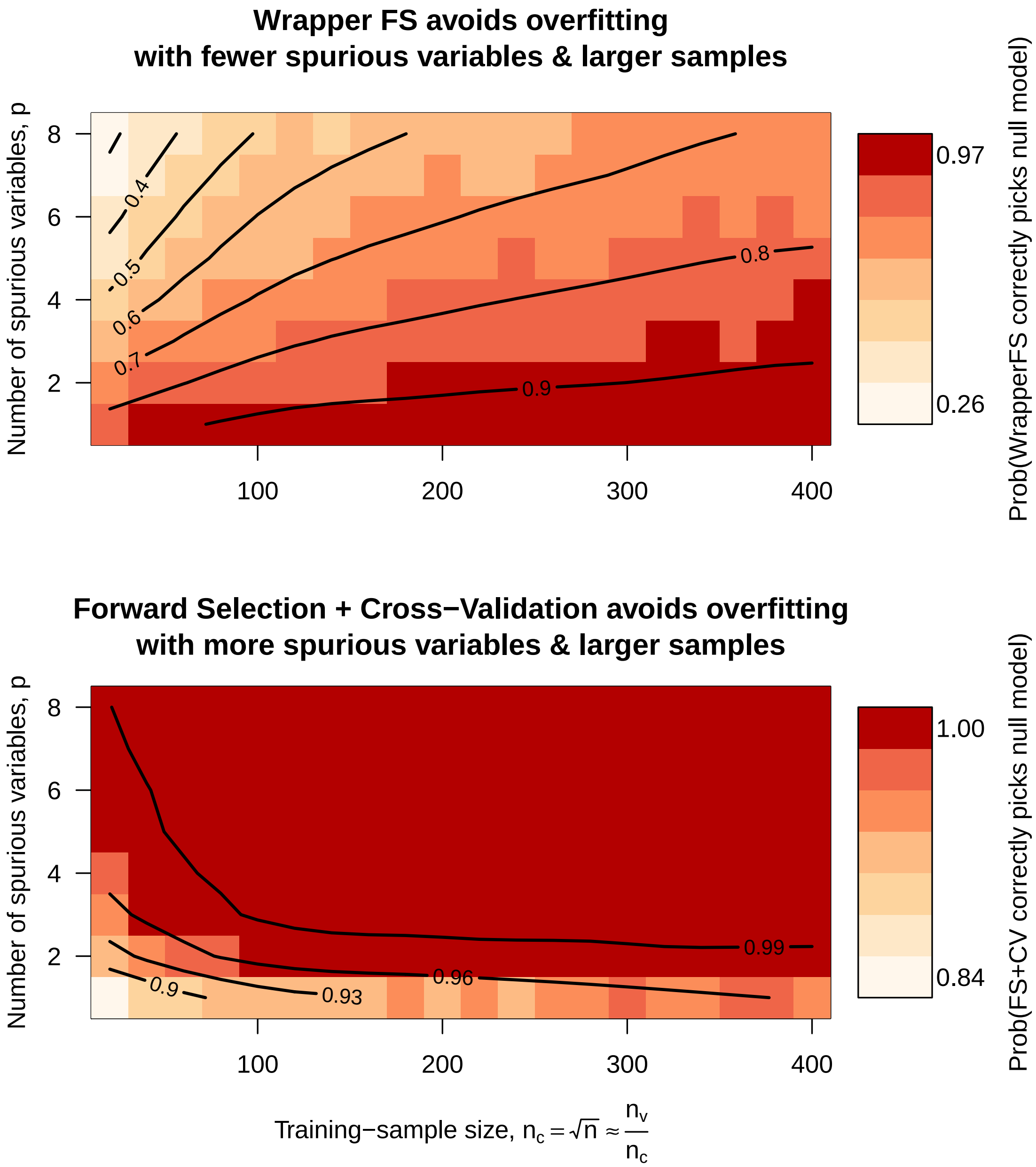

Figure 4.2: Top figure: $\mathbb{P}$(correctly choose null model) for Wrapper FS. The contours are generally shaped roughly like $p = \sqrt{n_c} \approx \sqrt{n_v/n_c}$. As per Theorem 4.4, Wrapper FS is model-selection consistent only if $p^2/n_c \to 0$ and $p^2 \cdot n_c/n_v \to 0$. Bottom figure: For FS+SeqCV, the contours are not at all like $p = \sqrt{n_c} \approx \sqrt{n_v/n_c}$. At every n_c and n_v/n_c, $\mathbb{P}$(correctly choose null model) rises with p, as we conjecture at the end of Section 4.2. Contour lines estimated from 2D loess fit, based on 500 simulations in each cell of the $p \times n_c$ grid.

model $J_h = J_* \cup h$:

$$B_h = \mathbb{E}_v \left(\widehat{MSE}(J_h) - \widehat{MSE}(J_*) \right)$$

where $\mathbb{E}_v$ is the expectation taken over validation datasets. Cross-validation makes a model-selection mistake if $\widehat{MSE}(J_h) < \widehat{MSE}(J_*)$, which depends on both the training and test datasets. We also speak of a "training mistake" if $B_h < 0$, which depends only on the training dataset: this is the event when an observed trained-estimate of the spurious model actually generalizes better than the trained-estimate of the true model.

Let $\mathbf{X}_*$ and Σ_* refer to only the first k covariates in $\mathbf{X}$, while $\Sigma_{J_h} = \begin{bmatrix} \Sigma_* & \Sigma_{*,h} \\ \Sigma_{*,h}^T & 1 \end{bmatrix}$ is the population covariance matrix for all covariates in J_* along with covariate h. Let X_h be just the column for covariate h alone. When we compare the true model against the spurious model $Y = \beta_1 X_1 + \ldots + \beta_k X_k + \beta_h X_h + \epsilon$, we will see that B_h has the form

$$\begin{aligned} B_h &= (\hat{\beta}_{J_h} - \beta)^T \Sigma (\hat{\beta}_{J_h} - \beta) - (\hat{\beta}_{J_*} - \beta)^T \Sigma (\hat{\beta}_{J_*} - \beta) \\ &= \tilde{\beta}_{J_h}^2 \cdot \left(\gamma_{J_h} + (\hat{\alpha}_{X_h} - \alpha_{X_h})^T \Sigma_* (\hat{\alpha}_{X_h} - \alpha_{X_h}) \right) - 2\tilde{\beta}_{J_h} \cdot \hat{\alpha}_\epsilon^T \Sigma_* (\hat{\alpha}_{X_h} - \alpha_{X_h}) \end{aligned}$$

for some $\gamma_{J_h} \in (0,1)$, where $\tilde{\beta}_{J_h} = \frac{X_{c,h}^T P_*^\perp \epsilon_c}{X_{c,h}^T P_*^\perp X_{c,h}}$ and $P_*^\perp = I - \mathbf{X}_{c,*}(\mathbf{X}_{c,*}^T \mathbf{X}_{c,*})^{-1} \mathbf{X}_{c,*}^T$; and $\hat{\alpha}_{X_h} = (\mathbf{X}_{c,*}^T \mathbf{X}_{c,*})^{-1} \mathbf{X}_{c,*}^T X_{c,h}$ estimates $\alpha_{X_h} = \Sigma_*^{-1} \Sigma_{*,h}$; and $\hat{\alpha}_\epsilon = (\mathbf{X}_{c,*}^T \mathbf{X}_{c,*})^{-1} \mathbf{X}_{c,*}^T \epsilon_c$.

Proposition 4.6. *Assume 1, 2, and 6. Consider comparing the true model against the* $p-k$ *spurious models* $Y = \beta_1 X_1 + \ldots + \beta_k X_k + \beta_h X_h + \epsilon$, *for candidate covariate* $h \in k+1, \ldots, p$.

Then, $\exists\ c > 0$ *such that, for* n_c *large enough, the probability of a "training mistake" vanishes as* $n \to \infty$*:*

$$\mathbb{P}_c(\min_h B_h < 0) \leq c \left(kp \sqrt{\frac{\log(p)}{n_c}} + p^{-1} \right) \to 0\,.$$

Proposition 4.7. *Assume 1, 2, and 6. Consider comparing the true model against the* $p-k$ *spurious models* $Y = \beta_1 X_1 + \ldots + \beta_k X_k + \beta_h X_h + \epsilon$, *for candidate covariate* $h \in k+1, \ldots, p$.

Then, $\exists\ c > 0$ *such that, for* n_c *large enough, the probability of a model-selection mistake vanishes as* $n \to \infty$*:*

$$\mathbb{P}\left(\min_h \widehat{MSE}(J_h) < \widehat{MSE}(J_*) \right) \leq c \left(kp \sqrt{\frac{\log(p)}{n_c}} + \frac{kp \log(p)}{\sqrt{n_v}} + \sqrt{\frac{n_c k p^2 \log(p)}{n_v}} + p^{-1} \right) \to 0\,.$$

4.3.3 $p^2/n_c \to 0$ is necessary for $\mathbb{P}(overfit) \to 0$

Under Assumptions 7 and 8, for a given spurious covariate h and a given training sample $(Y_c, X_{c,h})$, we will see that the expected difference in test errors is

$$B_h \equiv \mathbb{E}_v \left(\widehat{MSE}(J_h) - \widehat{MSE}(J_*) \right) = \hat{\beta}_h^2 (1 + \overline{X}_{c,h}^2) - 2\hat{\beta}_h \overline{X}_{c,h} \bar{\epsilon}_c$$

where $\mathbb{E}_v$ is the expectation taken over validation datasets. (This is a special case of the same B_h as in Section 4.3.2.)

Proposition 4.8. *Assume 7, 8, 9.*

Then $\liminf_{n\to\infty} \mathbb{P}_c(\min_h B_h < 0) \geq 0.12 > 0$, *where* $\mathbb{P}_c$ *is the probability taken over construction datasets.*

In other words, the probability of a training mistake (where at least one estimated model with spurious structure happens to generalize better than the estimated model with true structure) does not vanish.

Chapter 5

Practical choice of split ratio at large sample sizes

5.1 Yang's "CV paradox"

If n_c/n_v does not go to 0, then our probability of correct model selection is bounded away from 1. In fact, having more data can actually harm our model-selection performance. Yang (2006) calls this counterintuitive effect the "CV paradox." To supplement Yang's illustration of this effect on a classification problem, our simulations in Section 6.1 show that the same effect can arise with regression model selection.

Consider the top-left subplot of Figure 6.1. As a baseline, take the case of $n = 1250$ and 5-fold CV, where the probability of correct model selection is around 0.83. If we quintuple the sample size to $n = 6250$ and stay with 5-fold CV, the estimated success probability drops slightly to 0.82. Adding new data helps substantially only if it is mostly allocated to testing, as for inverse 5-fold CV, whose estimated success probability rises to 1.00 at $n = 6250$. Similar effects are seen throughout Figure 6.1.

In other words, the standard advice to use 5-fold or 10-fold CV (Breiman and Spector, 1992; Kohavi, 1995) may be adequate for prediction but not necessarily for model selection as n grows. However, in the next Section we show why practical finite-sample guidance is rarely possible unless some nontrivial knowledge about the unknown parameters is available.

5.2 Heuristic choice of n_c/n_v

Traditional high training ratios, 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, so that it may be safe to reduce n_c/n_v 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 MCCV, 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 overfit, since the probability of underfit goes to 0 regardless of n_c/n.

However, those probabilities are for Full CV. With Sequential CV 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 overfit evaluated at $p-k=1$,

$$\mathbb{P}\left(\hat{k} > k\right) \approx 1 - \mathbb{P}\left(\chi^2_1 < \left(1 + \frac{n}{n_c}\right)\right).$$

In the neighborhood of $n_c/n \leq 1/10$, 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 underfit, we refine an intermediate step in Zhang's proof of his Theorem 1, which derives relevant expressions for $\widehat{MSE}(J_t)$.

Corollary 5.1. *Let $\epsilon \sim N(0, \sigma^2)$ and let $\mathbf{X}$ be a fixed sequence in n. Assume the model path is fixed and the predictors are ordered, so that model J_h corresponds to using the first h predictors. Assume*

A′. $n_v \to \infty$ and $n_v/n = \lambda + o(1)$ where $\lambda \to 1$ as $n \to \infty$;

B. $\sup_{n_v \to \infty} \sup_s \|n_v^{-1} \mathbf{X}^T_{s,J_h} \mathbf{X}_{s,J_h} - V_h\| = o(1)$, where $V_h, h \leq p$ is a sequence of positive definite matrices, and $\sup_s$ is taken over all subsets of $\{1, \ldots, n\}$ of size n_v;

C′. For $h < k$, $b_{J_h} = \liminf_{n\to\infty} n^{-1} (\mathbf{X}\beta)^T P^{\perp}_{J_h} \mathbf{X}\beta > 0$ and $c_{J_h} = \limsup_{n\to\infty} n^{-1} (\mathbf{X}\beta)^T P_{J_h} \mathbf{X}\beta < c$ for some $c < \infty$;

D. For $h \leq p$, $\max_{i \leq n} H^{(h)}_{ii} \to 0$, where $H^{(h)}_{ii}$ are the diagonal elements of P_{J_h}.

Consider comparing true model J_* (of size k) against a particular underfitting model J_h (of size h), where $J_h \subsetneq J_*$. Then we have

$$\mathbb{P}(\text{correctly choose } J_* \text{ over } J_h) = \mathbb{P}\left(A_1 > 2A_2 + k\left(1 + \frac{n}{n_c}\right) - \frac{b_{J_h}}{\sigma^2/n} + o_p(1)\right)$$

where $A_1 \sim \chi^2_{k-h}$; $A_2 \sim N(0, b_{J_h} n \sigma^{-2})$; and A_1 and A_2 are not independent.

If we additionally assume that the $o_p(1)$ term is negligible for n larger than some sufficiently large N, then we can ensure the probability above to be at least $1-\alpha$ by choosing $n > N$ and training ratio at least

$$\frac{n_c}{n} \geq \left(\frac{\left(\sqrt{\frac{b_{J_h}}{\sigma^2/n}} - Z_{1-\frac{\alpha}{2}}\right)^2 + \chi^2_{(k-h),\alpha/2} - Z^2_{1-\frac{\alpha}{2}}}{k} - 1 \right)^{-1} . \tag{5.1}$$

Derivations are in Chapter 8. If n is large enough and the $o_p(1)$ term is negligible, our argument is conservative in giving the probability of avoiding underfit along a fixed path.

This Corollary only considers comparing J_* against a single underfitting model J_h. However, since b_{J_h} and $\chi^2_{(k-h),\alpha/2}$ are both monotonically nonincreasing along the model path (as h rises toward k), the worst case should be the model with $h = k-1$. If we choose n large enough to achieve high success probability on this worst case, the smaller models on that path should also have high individual probabilities (of successfully being rejected for $*$). We can control overall probability of underfit with a Bonferroni adjustment of α to α/k.

Choosing n_c/n We argued above that a training ratio around $n_c/n = 1/10$ is more than adequate to avoid overfit, whereas 10-fold CV's $n_c/n = 9/10$ is commonly used to avoid underfit. We might expect Corollary 5.1 to help us tune n_c/n and balance these competing tendencies. However, Equation 5.1 shows that this is impractical unless we have fairly good knowledge about b_{J_h} 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 5.1 then implies we need $\sqrt{1+\frac{n}{n_c}} \leq \sqrt{\frac{nb_{J_h}}{\sigma^2 k}}$, 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 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 6.2.2:

- If n is large and we confidently believe $\sqrt{\frac{nb_{J_h}}{\sigma^2 k}} \gtrsim \sqrt{1+\frac{10}{1}} \approx 3.32$, we can safely use a low training ratio of $n_c/n = 1/10$, avoiding both under- and overfit.

- Otherwise—if n is not large, or our initial guess of $\sqrt{\frac{nb_{J_h}}{\sigma^2 k}}$ is too small or imprecise—a conventional split ratio such as 8/10 or 9/10 (5-fold or 10-fold CV) is safer. We will be prone to overfit but at least ought to avoid underfit.

Tradeoffs between underfit and overfit Figure 5.1 illustrates the competing curves of $\mathbb{P}(\text{underfit})$ and $\mathbb{P}(\text{overfit})$ vs. n_c/n at various values of the signal-to-noise ratio $\frac{b_{J_h}}{\sigma^2/n}$ and true model size k. Across all n_c/n

ratios, the estimated probability of underfit (solid lines) become smaller as the SNR increases, but larger k makes underfit much more likely.

Finally, in Figure 5.2 we simulate empirical estimates corresponding to the curves from Figure 5.1. Our simulations used an orthogonal design with $n = 500$, $p = k + 1$, $\sigma^2 = 1$, and β's equal-valued nonzero coefficients set to achieve the target SNRs. We conduct 600 simulations at each value of n_c/n and SNR. For each dataset and training ratio, we estimate test MSEs along a fixed model path using Monte Carlo CV with 20 repetitions. Because we found the estimates from Equation 5.1 to be quite conservative, we illustrate a range of smaller SNRs here.

- $\widehat{\mathbb{P}}(\text{underfit}) = \widehat{\mathbb{P}}\left(\exists\, h < k : \widehat{MSE}(J_h) < \widehat{MSE}(J_k)\right)$. The simulations indicate that our heuristic underfit curves from Equation 5.1 are conservative, but the approximate patterns are the same, as the probability of underfit decreases with training ratio. $\widehat{\mathbb{P}}(\text{underfit})$ appears to rise very slightly with k, as expected from Equation 5.1.
- $\widehat{\mathbb{P}}(\text{overfit}) = \widehat{\mathbb{P}}\left(\widehat{MSE}(J_{k+1}) < \widehat{MSE}(J_k)\right)$. Zhang's asymptotic probability of overfit appears slightly anti-conservative by our simulations, especially at high training ratios. We estimate separate $\widehat{\mathbb{P}}(\text{overfit})$ curves for each SNR and plot all (dashed) lines. As expected, they overlap almost perfectly and are not affected much by SNR or k.

Both figures confirm that high training ratios avoid underfit, low ratios avoid overfit, and there is only a narrow range of $\sqrt{b_{k-1}}$ in which it makes sense to tune n_c/n.

Finally, it may also be impossible to make good assumptions about k, σ^2, and b_{m_h}. In this situation, instead of hoping to correctly select one model, we may prefer to explicitly acknowledge the uncertainty in the model selection process. In Chapter 7 we discuss ways to build confidence sets for model selection with cross-validation.

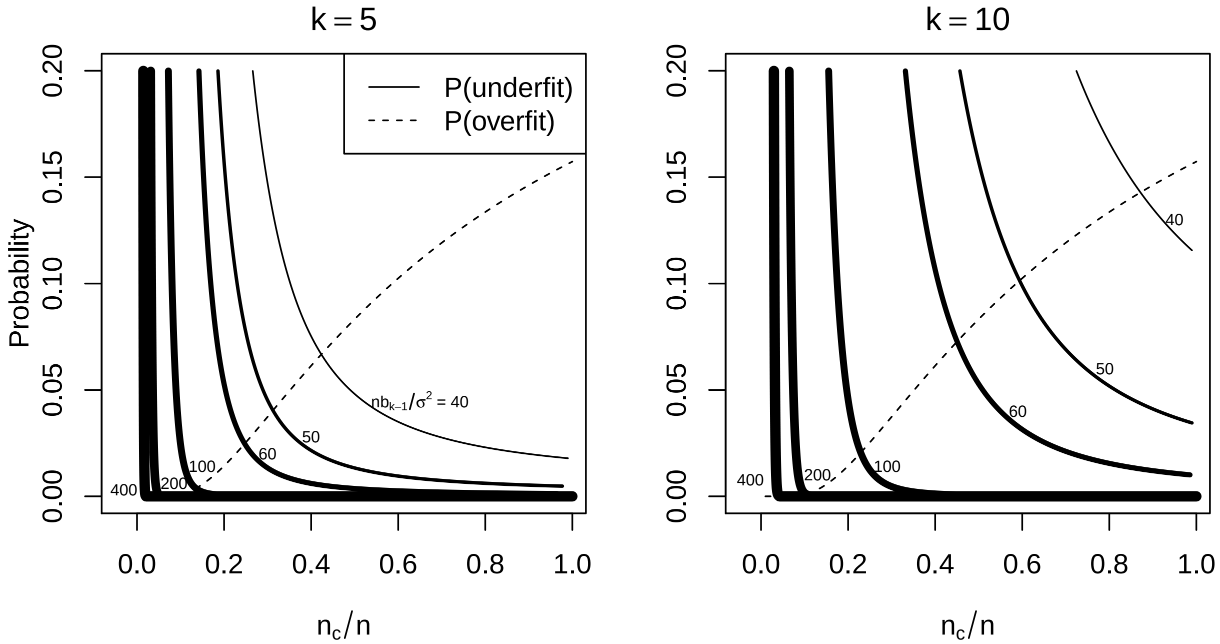

Figure 5.1: For $k = 5$ (left figure) and $k = 10$ (right figure), we plot the $\mathbb{P}$(overfit) (dashed line) and the $\mathbb{P}$(underfit) (solid lines at different levels of signal-to-noise ratio), vs. training ratio. Estimated using Zhang (1993) and Corollary 5.1.

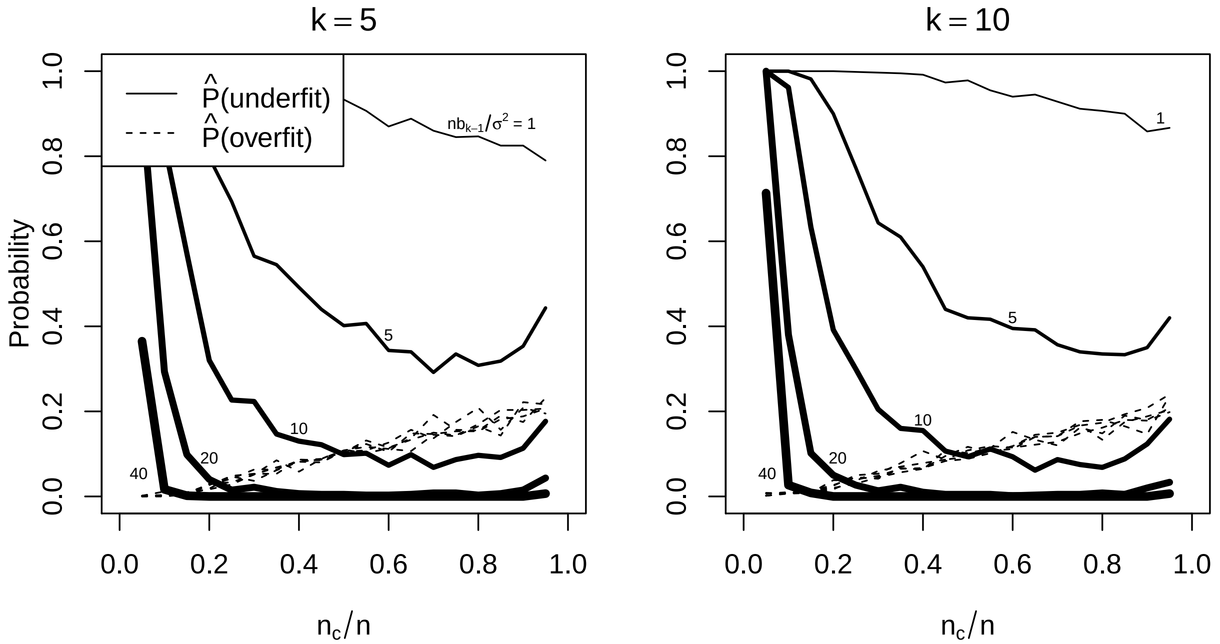

Figure 5.2: For $k = 5$ (left figure) and $k = 10$ (right figure), we plot the $\widehat{\mathbb{P}}$(overfit) (dashed lines) and $\widehat{\mathbb{P}}$(underfit) (solid lines), at different levels of signal-to-noise ratio, vs. training ratio. Estimated from 600 simulations at each combination of n_c/n and SNR, with orthogonal Gaussian design & noise, a fixed correct path, $n = 500$, and $p = k + 1$.

Chapter 6

Simulation studies and real-data examples for FS+SeqCV

We use large-scale simulations in Section 6.1 to empirically corroborate the theory from Chapters 3 and 4. We also demonstrate the performance of FS with SeqCV on several real datasets. Section 6.2.1 shows that FS+SeqCV performs as well as standard methods on two commonly-used benchmark datasets, despite fairly small sample sizes. In Section 6.2.2, we show that FS+SeqCV with a low train/test ratio outperforms competitors on the much larger Million Song Dataset of Bertin-Mahieux et al. (2011).

6.1 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: train on one fold and test on the remaining $V-1$ folds. We also contrast our Sequential CV vs. standard Full CV.

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 n range, together with a small $\beta_{min} = 0.2$, adequately contrasts the low-signal and high-signal cases. The k nonzero coefficients were drawn from a Uniform distribution, then shifted and scaled to have the range $[\beta_{min} = 0.2, \beta_{max} = 2]$. The design matrix $\mathbf{X}$ was drawn from a Normal distribution with 0 mean and covariance matrix Σ, using one of two correlation structures:

- 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 $\mu < (2k-1)^{-1}$.

- Setting 2: $\Sigma_2(\mu)$ has the following 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 is not always 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.

The noise ϵ was drawn independently from one of two distributions: either a sub-Gaussian $\epsilon \sim N(0,1)$, or a heavy-tailed $\epsilon \sim t(df = 2)$ which has no finite second moment. (The simulations with t_2 noise or with covariance structure Σ_2 use a range of smaller p and k values, which we found to adequately illustrate the difficulty of model selection at these settings.)

We run at least 400 replicate simulations at every combination of data-generation settings, independently generating new datasets and running every estimation procedure on each dataset. In Figures 6.1-6.5, error bars show $\pm 2 \cdot SE$ as approximate marginal 95% confidence intervals.

For Theorem 3.1, the black lines in Figure 6.1 illustrate the path-consistency of FS, plotting the probability of correct model selection for oracle FS with known k. As expected, the problem becomes easier in each sub-plot with increasing signal-to-noise ratio (SNR). Higher μ makes the problem a little harder but not impossible, as we see by comparing the left and right halves of the figure. As expected from our beta-min condition, higher p and higher k both make the problem harder, causing ever-higher values of n to count as "low SNR" conditions. (Note that larger k implies smaller $p - k$ so there are fewer ways to underfit, and our simulations also use smaller μ at larger k. However, these benefits of larger k do not appear to outweigh the harms, at least on the scale of k simulated here.)

For Theorem 4.1, Figure 6.1 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 underfit (despite low chance of overfit). However, the same low training-ratio performs best at high n, due to negligible chance of underfit when SNR is high. Inverted 5-fold CV approaches success probabilities of 1 at very high n, in all but the highest p, k, μ condition. Still, the values of n that correspond to "low SNR" get larger with higher p, k, and μ. Meanwhile, overfit 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 $n = 1250$. On the other hand, as we conjecture in Section 4.2, this plateau does get better at higher p, although the benefits of high p for CV happen only at high SNRs. When starting from moderate SNRs, raising p merely decreases the SNR and causes underfit instead of preventing overfit.

In addition to the effects of high vs. low training ratio, Figure 6.1 also contrasts between SeqCV vs. FullCV (dark vs. light colors). In several low-k cases, 5-fold FullCV (light) overfits more than 5-fold SeqCV (dark) 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. (Not shown: we also simulated single-split vs. V-fold CV, finding that that a single data-split tends to underperform V-fold CV at the same training ratio, as expected.)

Next, Figures 6.2 and 6.3 use the same layout as Figure 6.1, but instead report the average numbers of false negatives (underfitting) and false positives (overfitting) on the vertical axes. As expected, each algorithm generally improves with higher n but suffers at higher p, k, and μ. In both figures, oracle FS performs poorly at low SNRs because it can neither stop early (to avoid adding spurious variables) nor continue late (to collect all true variables after some spurious variables were added early). All of the CV methods tend to stop too early at low SNRs, having more false negatives and fewer false positives than oracle FS. Likewise, SeqCV has more false negatives but fewer false positives than FullCV, and inverted 5-fold CV has more false negatives but fewer false positives than regular 5-fold. However, the CV methods have fewer false positives with higher p, as we conjecture in Section 4.2, at least until the high-μ and $p = 1250$ case. Even this favorable Σ_1 correlation structure can suffer from high $\mu > (2k-1)^{-1}$ if p is too large.

Across the three figures, the transition between "low" and "high" SNR can differ for each algorithm. For instance, $n = 1250$ and $k = 125$ appears to be a borderline SNR at some p, μ: both SeqCV methods underfit dramatically, and inverted 5-fold FullCV overfits substantially, but regular 5-fold FullCV only overfits a little.

Returning to success probabilities, Figure 6.4 illustrates the effect of heavy-tailed noise, using a similar layout to Figure 6.1 but drawing super-Gaussian $\epsilon \sim t(df = 2)$ instead. Model-selection becomes uniformly more difficult when the noise has no finite second moment. Nonetheless, consistency is not ruled out: several of the subplots in Figure 6.4 do show success probabilities approaching 1 for oracle FS, and none of the FS+CV variants have plateaued yet at the largest sample size shown. In fact, the same simulations with t_3 noise rescaled to unit variance (not shown) look identical to the Gaussian noise, despite the heavy tails.

Finally, Figure 6.5 shows the effect of the "worst-case" correlation structure $\Sigma_2(\mu)$. This simulation setup was similar to selected subplots of Figure 6.1, 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 6.5. 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 Σ structures as seen in Figures 6.1 and 6.4.

Several other simulation settings (Toeplitz correlation matrix; deterministic β vector with decreasing nonzero entries; higher β_{max}; lower μ; repeated V-fold CV, MC CV, and CV-v variants) did not lead to

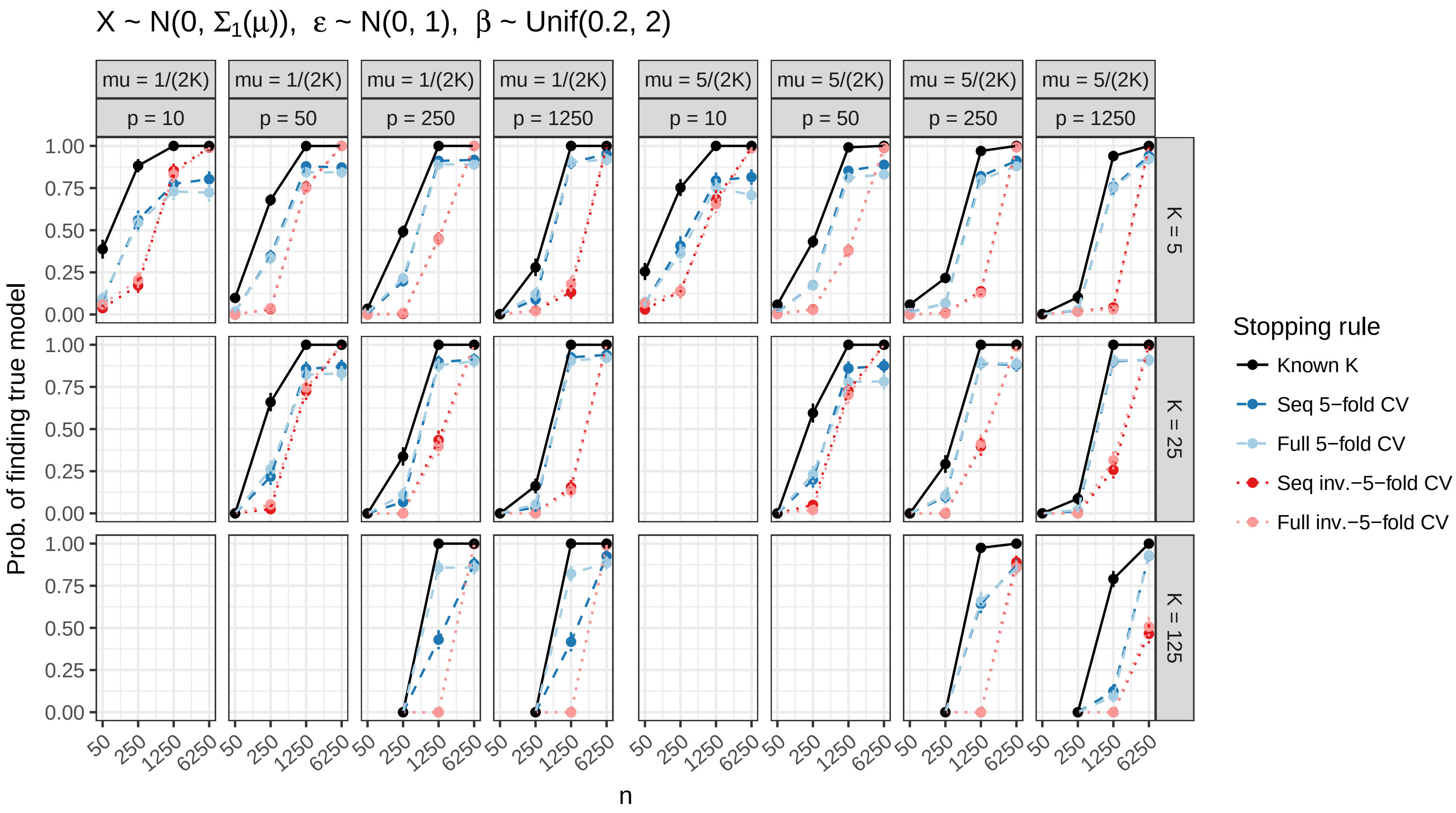

Figure 6.1: Simulations to illustrate selection-consistency of FS, as per Theorems 3.1 and 4.1. Oracle FS (solid black line) approaches success probability 1 as n increases in every setting, but the problem is harder at larger μ (left vs. right halves of plot), p, and k. High training ratios (5-fold CV, dashed line) underfit less at small n but overfit more at large n, while low training ratios (inverted 5-fold CV, dotted line) act the opposite. SeqCV (dark colors) and FullCV (light colors) are almost indistinguishable, except for a few high-k cases where 5-fold SeqCV stops too early more often than 5-fold FullCV does. Based on 400 simulations at each data point. Error bars show $\pm 2 \cdot SE$.

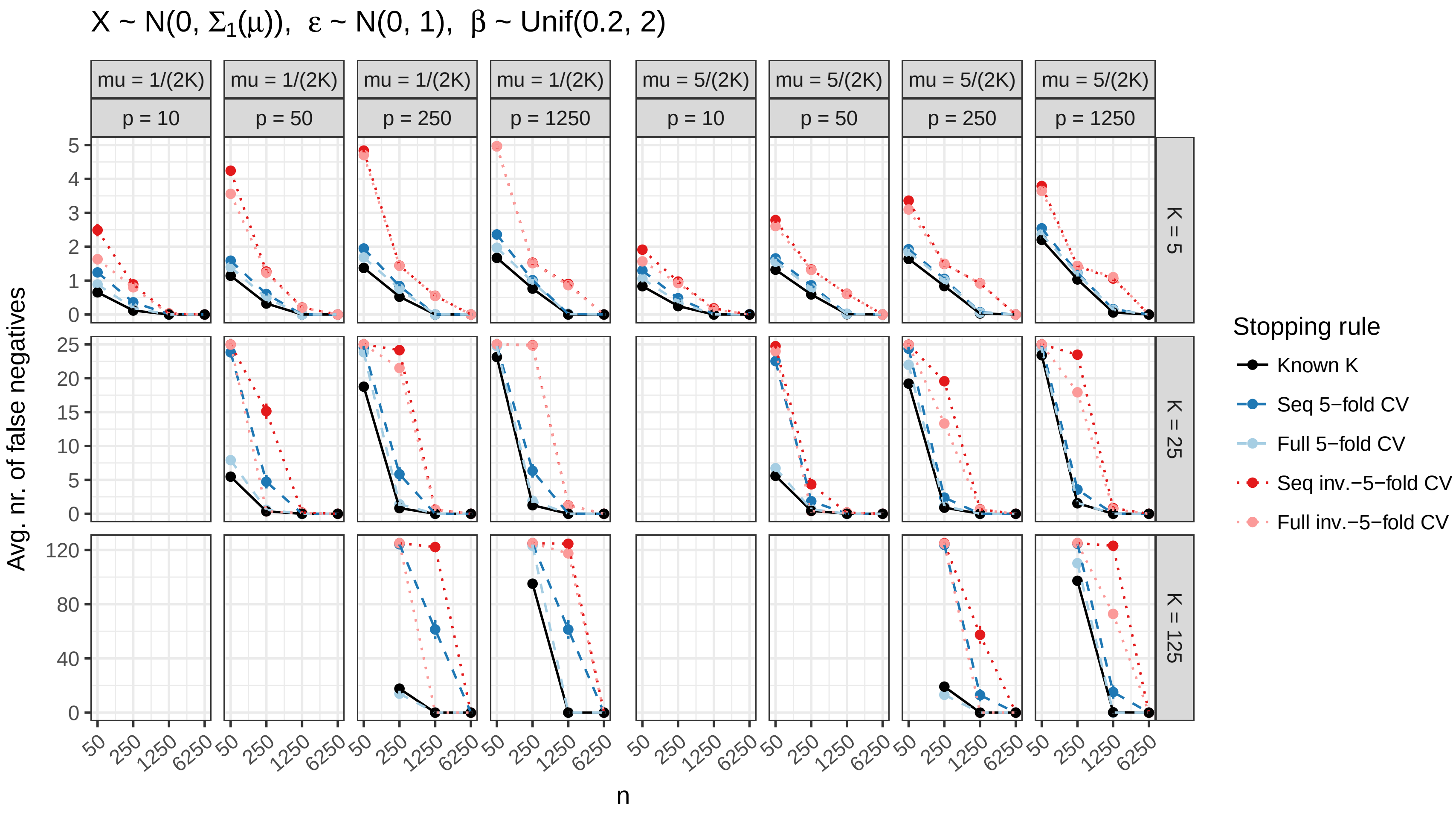

Figure 6.2: Simulated average counts of false negatives when using FS, illustrating the role of underfit in Theorems 3.1 and 4.1. Oracle FS (solid black line) approaches 0 false negatives as n increases in every setting, but the problem is harder at larger μ, p, and k, since oracle FS is not allowed to continue past k steps even if spurious variables were chosen early. High training ratios (5-fold CV, dashed line) underfit less than low training ratios (inverted 5-fold CV, dotted line) do, but both tend to underfit less as n increases. FullCV (light colors) tends to underfit less than SeqCV (dark colors). Based on 400 simulations at each data point. Error bars show $\pm 2 \cdot SE$.

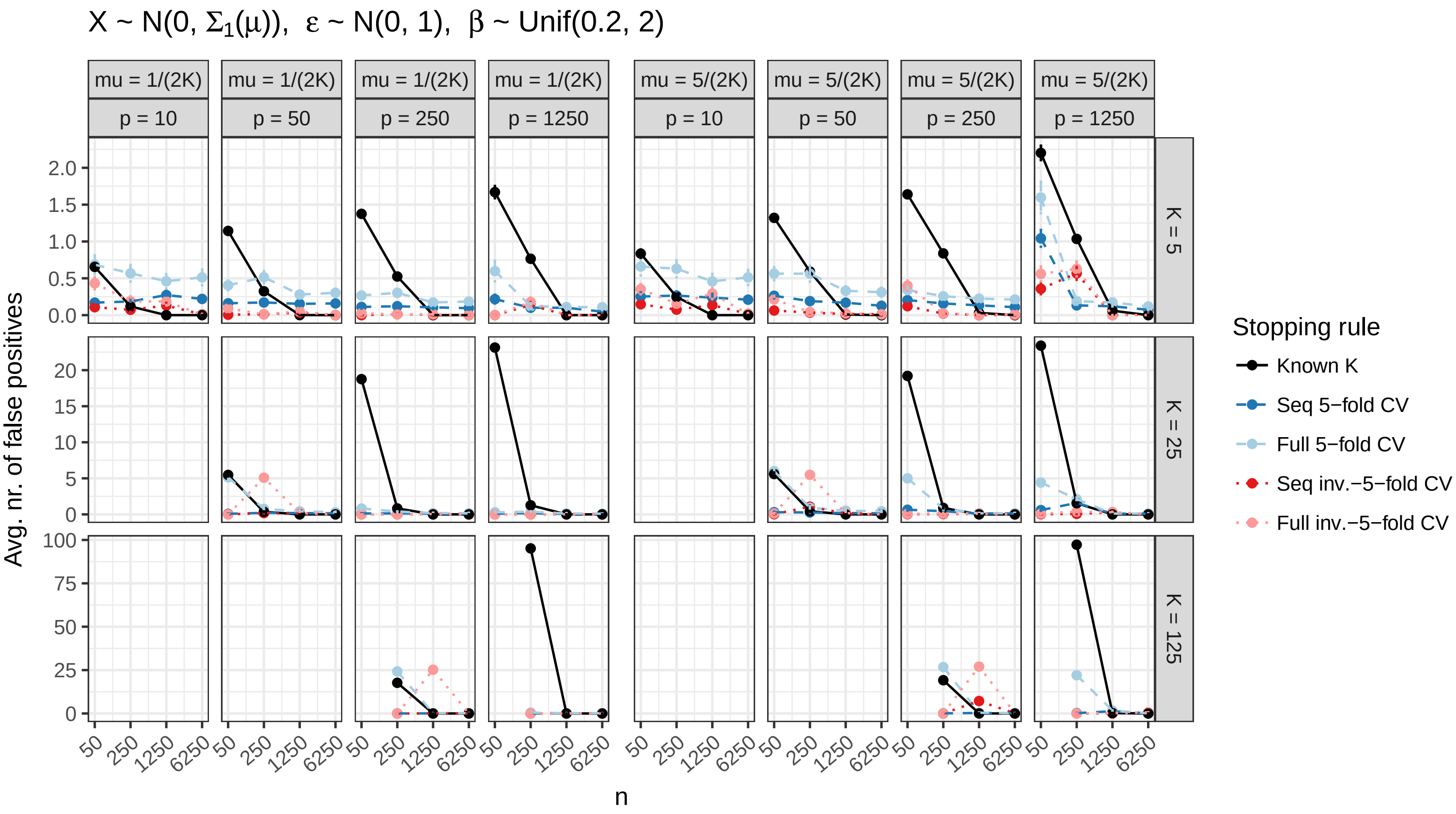

Figure 6.3: Simulated average counts of false positives when using FS, illustrating the role of overfit in Theorems 3.1 and 4.1. Oracle FS (solid black line) approaches 0 false positives as n increases in every setting, but the problem is harder at larger μ, p, and k, since oracle FS is not allowed to stop early even if it cannot detect the remaining true variables. High training ratios (5-fold CV, dashed line) overfit at every n, while low training ratios (inverted 5-fold CV, dotted line) tend to overfit less as n increases. SeqCV (dark colors) tends to overfit less than FullCV (light colors). Based on 400 simulations at each data point. Error bars show $\pm 2 \cdot SE$.

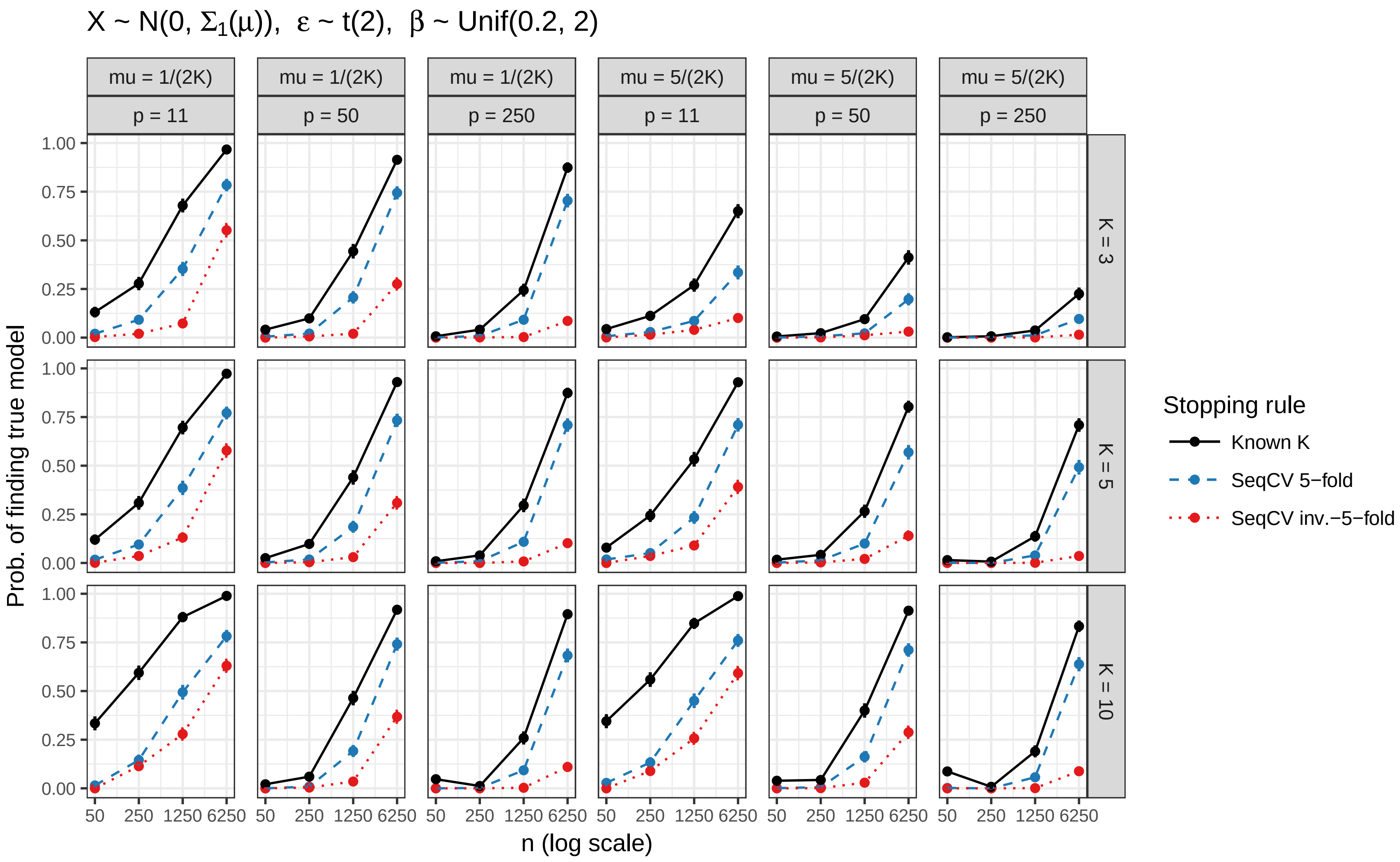

Figure 6.4: Simulations to illustrate failure of selection-consistency of FS under heavy tails. Similar setup as Figure 6.1 but with $t(df = 2)$ noise, which has no finite second moment. As before, the problem is easier for larger n, smaller μ, and smaller $p - k$. (On this limited range of k values, larger k actually helps, but note that our settings with larger k have smaller μ). Our theory does not address heavy tails and simulation results do not rule out consistency here, but achieving high success probability requires much higher n here than for light-tailed data. Low training ratios (dotted line) seem least robust to heavy tails. Based on 1000 simulations at each data point. Error bars show $\pm 2 \cdot SE$.

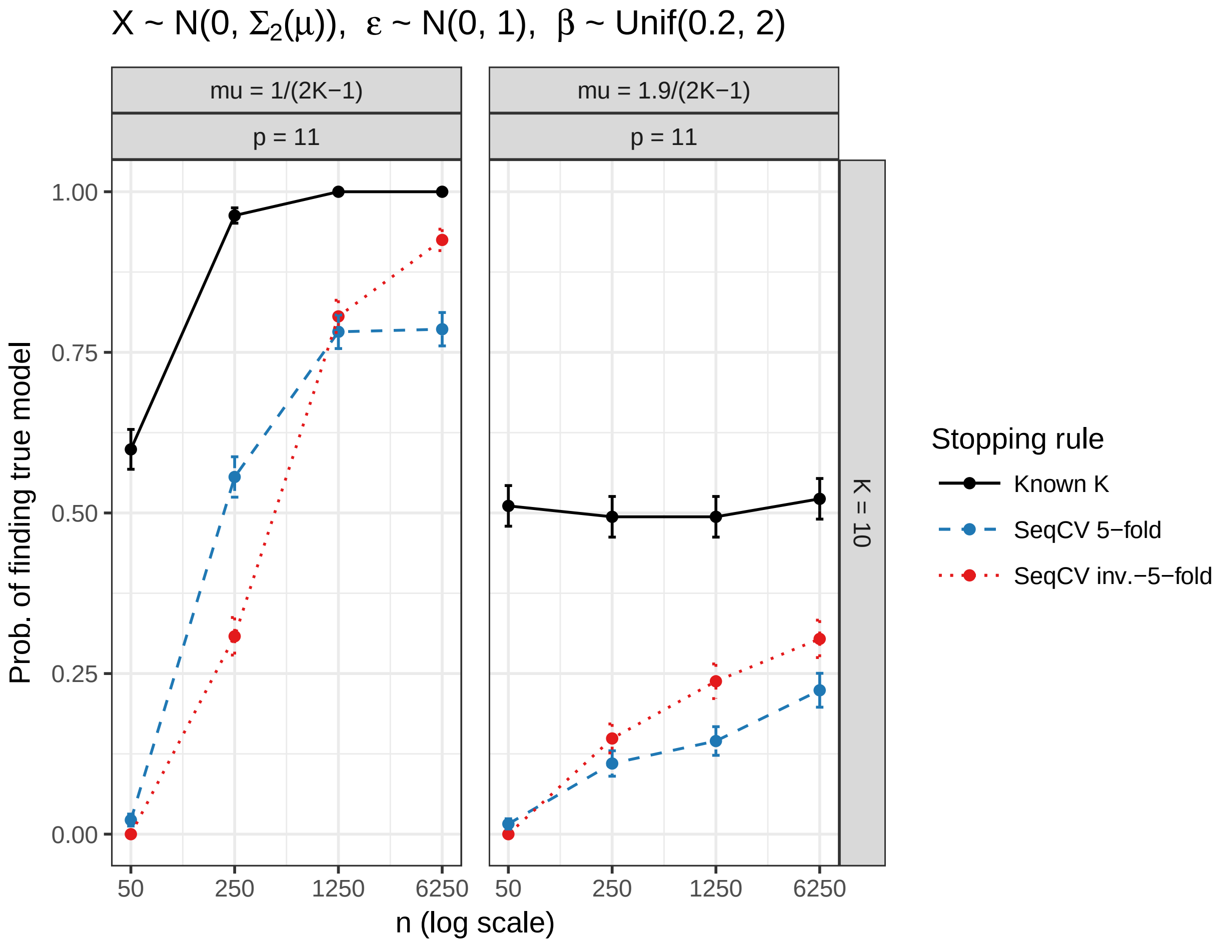

Figure 6.5: Simulations to illustrate failure of selection-consistency of FS under a "worst case" correlation structure. Similar setup as subplots of Figure 6.1, but with different levels of μ and using design structure $\Sigma_2(\mu)$, for which the coherence condition $\mu < (2k-1)^{-1}$ is necessary. When the condition is not met (right subplot), the probability of success remains around 0.5 for oracle FS (solid black line) and much lower for FS+CV variants, even as n rises. Based on 1000 simulations at each data point. Error bars show $\pm 2 \cdot SE$.

substantially different results. We did find that V-fold outperforms single-split CV, but our plots omit the unsurprising single-split results to avoid visual clutter.

Simulations were conducted in `R` (R Core Team, 2018), using the packages `leaps` to implement FS (Lumley and Miller, 2017), `doParallel` to run simulations in parallel (Revolution Analytics and Weston, 2015), and `ggplot2` to plot results (Wickham, 2009). `R` code to reproduce our simulations and data analyses is available online at: `https://github.com/civilstat/wieczorek-thesis-code`

6.2 Real-data examples

6.2.1 Benchmark datasets

We illustrate the use of FS+SeqCV on two classic datasets: the prostate cancer and Boston housing data. In replicating previous analyses, we find that FS+SeqCV performs as well as competing methods.

Prostate cancer data

Hastie et al. (2009) illustrate several approaches to linear regression selection or shrinkage on the classic prostate cancer dataset of Stamey et al. (1989). The task is to predict the logarithm of prostate-specific antigen levels, using $p = 8$ predictor variables including patient age and several clinical measures. Data are available at the patient level for $n = 97$ male patients. Hastie et al. have split the data into a learning set of 67 cases and a holdout set of 30 cases, so cross-validation is applied only to the 67 learning cases. Our Figure 6.6 mimics the top-left subplot in Hastie et al.'s Figure 3.7, which plotted estimated prediction error against model size, using 10-fold CV with All Subsets regression. With so few cases, we agree that 10-fold CV's high training ratio is appropriate.

The model paths chosen by FS and by All Subsets regression tend to match, and our CV error estimates match Hastie et al.'s Figure 3.7 using FS instead of All Subsets. Furthermore, they do not use FullCV, but rather the "1 standard error rule." They choose the smallest model whose mean CV error was within 1 SE of the global minimizer's mean CV error. Although FullCV would keep 7 of the 8 variables in the model, the 1SE rule chooses a model size of only 2 variables, which happens to coincide with SeqCV.

Finally, we fit a FS path to the full learning set, stop at the chosen subset size $\hat{k}$, and evaluate this model's predictions on the holdout set. The holdout error for the full OLS model ($\hat{k} = p = 8$) is 0.521, and FullCV ($\hat{k} = 7$) has holdout error of 0.517, while the holdout error for the FS+SeqCV model ($\hat{k} = 2$, also the AllSubsets+1SE model) is only 0.492. In other words, for 10-fold CV, FS+SeqCV performs identically to All Subsets with the 1SE rule on the classic prostate dataset. By holdout error, FS+SeqCV outperforms FullCV or the full OLS model.

Boston housing data

Zhang (2011) evalutes OMP and his own proposed "FoBa" algorithm on the Boston housing dataset of Harrison and Rubinfeld (1978). The task is to predict median housing value (in thousands of dollars) from $p = 13$ other variables related to housing, pollution, and demographic and economic measures. Data are measured at the Census tract level, using $n = 506$ Boston-area tracts from the 1970 Census. In our Figure 6.6, we mimic Zhang's Figure 5, which plots training and test estimates of prediction error against model size.

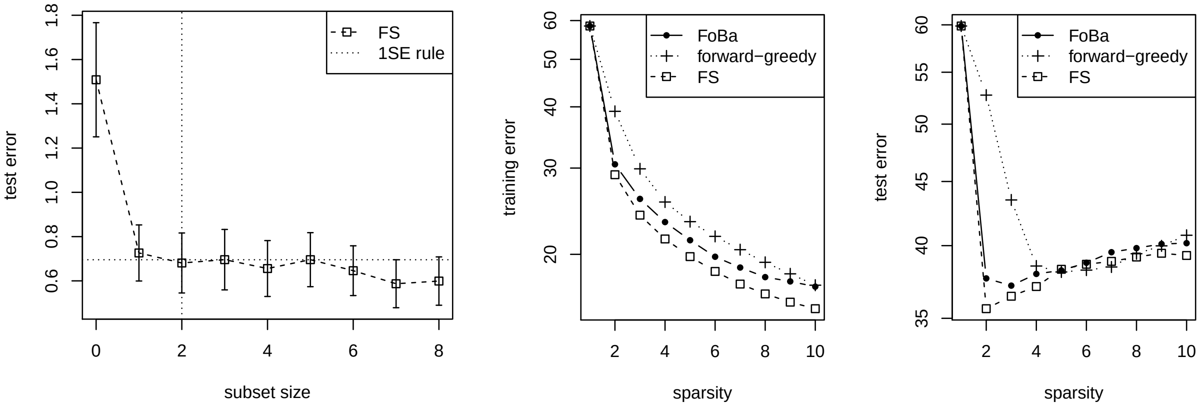

Figure 6.6: Left figure: Our replication of Figure 3.7 of Hastie et al. (2009), but replacing All Subsets with FS. We use their learning subset (67 cases) of the prostate cancer dataset and make one random partition into 10 folds. For each fold, train a full model path using FS on the data outside that fold, and record the test MSE at each subset size. The plots show average MSEs and their standard errors over the 10 folds. Right two figures: Our replication of Figure 5 of Zhang (2011), but replacing the Lasso with FS. We partition the Boston housing dataset into splits with $n_c = 50$ and $n_v = 456$; train a full model path using each of FS, FoBa, and "forward-greedy" (OMP); record their training and test MSEs at each sparsity level; and average both MSEs over 500 random partitions.

We explicitly compare FS to Zhang's FoBa and to OMP (which he calls "forward-greedy"). FoBa is a special forward-backward stepwise variant of OMP that is allowed to take many backward steps (if needed) after every forward step. We use Zhang's R implementation of FoBa and OMP, available online at:
`http://tongzhang-ml.org/software.html`

Following Zhang, we do not set aside a holdout set, but simply use repeated MC CV with a training ratio of $n_c/n = 1/10$ on the entire sample. We repeatedly partition the data 500 times into 50 training and 456 testing cases; train a full model path using each algorithm; and report training and test estimates of MSE for each sparsity level from 1 to 10. Finally, to be comparable with Zhang's results, we do not include an intercept in the model by default, but treat it as a separate feature which may be added (or dropped, by FoBa) at any time.

For each algorithm, the CV test error curve has only one local minimum, so SeqCV and FullCV both choose the same sparsity levels. FS chooses a model with 2 predictors, compared to FoBa's choice of 3 predictors and OMP's choice of 5. Of the three models, FS is sparsest and has lowest test error.

Conclusions from benchmarks

For both of these exemplar datasets, there is no substantial difference in performance between FS+SeqCV and other methods used in common practice. Of course, examples can exist where a small early uptick in estimated CV error causes FS+SeqCV to choose a model which is far too small—but these are most probable in low-signal settings, where correct model-selection is hopeless for any algorithm.

At the other extreme, when both n and p are huge, SeqCV can provide substantial savings in computational cost or runtime compared to standard FullCV. In these settings, we can also cut costs and improve performance by choosing a low training ratio, as the next example demonstrates.

6.2.2 Million Song Dataset

We illustrate FS+SeqCV on a large dataset, where a small train/test ratio 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 Jehan (2010), the documentation for The Echo Nest "Analyze" API used to preprocess the MSD, "*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 (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 9.5 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 6.7 shows several high correlations between pairs of predictors, but not many. Although we do not meet our theorems' coherence threshold, we are not too concerned in light of the simulation results

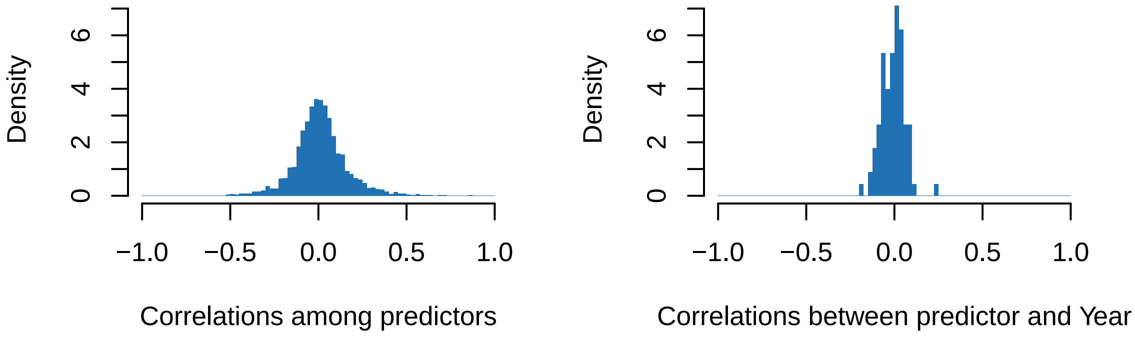

Figure 6.7: Histograms of correlations in the MSD dataset.

in Section 6.1. Also, the condition number of our predictors' correlation matrix is around 13.3, below the conventional multicollinearity cutoff of 30.

Model selection and evaluation

We illustrate the use of our proposed method, FS+SeqCV with a low train:test 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.

Following our heuristic advice in Section 5.2, we believe that a 10:90 split is reasonable here. The learning 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 $\sqrt{\frac{nb_{J_h}}{\sigma^2 k}} \gtrsim \sqrt{1 + \frac{10}{1}} \approx 3.32$. 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 $\sqrt{\frac{nb_{J_h}}{\sigma^2 k}} \approx \sqrt{\frac{463{,}715}{30}} \times \frac{0.3}{9.6} \approx 3.9 > 3.32$, 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 train:test ratio, as well as FS+FullCV at both train:test 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 each CV-based stopping rule and split ratio, we fit a model path on the first n_c learning cases, then use the next n_v learning cases 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.

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 ($R^2 = 0.24$). This difference in RMSEs translates to 1 year and 4 months, so even the full linear model does not improve predictions dramatically

Stopping rule	$\hat{k}$	RMSE (years)	Time (minutes)
Null model	0	10.85	<1
FS+SeqCV, 10:90	23	9.61	1
FS+SeqCV, 90:10	29	9.57	14
FS+FullCV, 10:90	60	9.52	10
FS+FullCV, 90:10	76	9.51	93
Full model	90	9.51	<1

Table 6.1: Selected model sizes, holdout RMSEs, and computation times for the MSD, under different stopping rules.

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.

For each approach, Table 6.1 reports the size of the selected model, holdout RMSE estimates (in years), and computation time (in minutes). As we move down the table's rows, we modestly reduce RMSE but dramatically increase model size and computation time.

First, we note in Table 6.1 that a lower training ratio (10:90 vs. 90:10) does not substantially change the selected model's RMSE, but it does 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 difference, especially with data recorded to the nearest year.

In the sparsest case, FS+SeqCV at the 10:90 split ratio selects a model with 23 variables, around a quarter of the original 90 predictors. This is a considerable reduction in model size and computational resources required, with negligible effect on predictive performance.

There is also considerable overlap among the predictors selected by the sparse models. For instance, both the Means and Variances of timbre coefficients 1, 2, 6, and 11 are selected in every sparse model, so these four basis functions appear to be among the most informative summaries of a song's content as it relates to Year.

Table 6.1 also reports the approximate runtime for each selection algorithm. Using a small 10:90 split is substantially faster than a large 90:10 split, because the computationally-expensive training is run on far less data. Likewise, using SeqCV 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 with the largest model in our scope, but which requires far fewer predictor variables and speeds up computation considerably. 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.

Chapter 7

Conclusions

7.1 Summary and 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 Sequential CV. We have also argued for the benefits of using a low training ratio for CV when conditions are suitable.

However, previous authors such as Harrell (2015) have argued that automatic regression model selection, such as FS, is almost never a good idea. Not only do the naive estimated inferences (p-values, CIs, etc.) fail to account for the selection process, but the choice of model itself is noisy and less interpretable than it appears. They point to decades of literature with conflicting advice as another reason not to trust such methods.

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. In this setting, we recommend using subject-matter expertise instead of selection algorithms. Dawes (1979) goes even further, arguing that in these settings we should not even try to estimate a linear model. Instead, devise an "improper linear model" which simply assigns equal weights to each predictor that experts agree is "relevant," with signs also chosen by the experts; and report its performance on the data.

When, then, are methods like FS valuable? One modern use-case for FS could be A/B testing for software products with large user bases. Websites or apps often want to evaluate proposed changes to the user interface (UI), testing the change on a small proportion of their audience before rolling out the change to all users. They may also be interested in understanding different effects for different subgroups of users, so an interpretable linear model could have value over a pure prediction model. For instance, maybe the proposed UI update hurts engagement among long-standing users, but helps among newer users.

With millions of registered users to choose from, and the ability to block or stratify carefully before randomizing users to the A or B arms of the experiment (current UI vs. updated UI), it can be possible to achieve a nearly-orthogonal design with high signal-to-noise ratio. Local UI expertise and data from previous A/B tests can help estimate reasonable values of k, σ^2, and b_{min} and even establish rough power calculations, based on the heuristics in Section 5.2. If so, we can actually expect FS to perform well at sifting through the covariates on each user (e.g., device and operating system; approximate geographic location; how long they have been a registered user, and other usage patterns; perhaps demographic information, if users fill out profiles; etc.) and selecting a good sparse linear model. This is also an instance of the large-n situation where our two suggestions for reducing computation (SeqCV rather than FullCV; and low rather than high train:test ratios) are expected to improve our chance of correct model-selection.

7.2 Future work

A different perspective on model selection with cross-validation, not specific to Forward Selection or other greedy path algorithms, is the idea of building a confidence set of models. Lei (2017) introduced the idea of "cross-validation with confidence" (CVC), using the models compared by cross-validation to build a $100(1-\alpha)\%$ confidence set which contains the best model with probability at least $1-\alpha$. This concept has several benefits over selecting a single model with CV. By building such confidence sets, data analysts can express the uncertainty in the model-selection process. Also, choosing α allows statisticians to trade off higher coverage vs. larger confidence set sizes in a familiar way, without tuning the CV training ratio. Finally, subject matter experts can choose one of the models in the confidence set based on their own criteria. Alternately, an automated algorithm can follow the "bet on sparsity" principle and choose the smallest model in the set.

Lei's framework is useful and widely applicable, with a proposed methodology assuming only that the loss function evaluated on individual test cases follows some distribution with sub-exponential tails. Lei's simulations also show that CVC's coverage tends to be very close to the nominal confidence level. However, the implementation calls for a bootstrap multiplier approach, which can add substantial computational expense on top of CV as well as a non-trivial layer of complexity to the code.

In future work, we propose to seek simpler and cheaper approximations to CVC that still have reasonable performance guarantees, as illustrated in the tentative results below. We suggest several approximations to CVC inspired by different approaches to "ranking with confidence," and we demonstrate their empirical performance in simulations. The influential paper of Dietterich (1998) also proposed and compared several approximate statistical tests for comparing two fitted models, including CV-based approaches. However, the justifications were heuristic and only applied to comparing two models, not to constructing a confidence set for the best of many models.

Additionally, in future work we will explore integrating Lei's CVC into FS+SeqCV using a sequential testing approach: stop adding variables when the next-larger model is not significantly better than the current model. Under the conditions for path-consistency of FS, we do not need to worry about information leaking from the test data back to the training data except on a set of measure approaching 0, until we reach the correct model. Hence, with sufficient data we should not stop too soon, and we should approach the nominal coverage level for our test of whether to continue past the true model. Unlike the discredited traditional significance-test stopping rules for FS, this "FS+SeqCVC" would account for the fact that we have been testing more than one pair of models.

Ranking with confidence Klein et al. (2018) propose a simple method to construct confidence sets for ranking populations, requiring only a multiple-comparisons-corrected confidence interval for each population's point estimate. The rectangular joint confidence region contains all estimates simultaneously with the desired probability, and the full procedure of Klein et al. gives a joint confidence set for all of the ranks. However, if we are only interested in the top rank, the procedure simplifies to selecting every population whose CI overlaps the CI for the top-ranked population. Although quite conservative, this method is valid whenever the individual CIs are valid.

Inspired by Klein et al. (2018), we propose a simple method "KWW" (Klein-Wright-Wieczorek) for CV with confidence. Say that we are comparing M different models using CV. For each model J being compared, estimate its cross-validation test-set $\widehat{MSE}(J)$ and the MSE's standard error $\widehat{SE}(J)$. Build a Bonferroni-corrected 2-sided Gaussian confidence interval (CI) for each of the M models: $\widehat{MSE}(J) \pm Z_{\alpha/(2M)}\widehat{SE}(J)$. Finally, let our confidence set contain the winning model and any model whose CI overlaps the winner's CI.

In a different approach, Hung and Fithian (2016) explore the problem of "rank verification." They study the naive procedure of testing whether the highest-ranked entity is significantly different from the runner-up, and so on down the ranks, stopping when a non-significant difference is found. They argue that for exponential families, an unadjusted two-tailed pairwise test comparing the estimated winner and runner-up is in fact a valid level-α test (and likewise for the next tests down the line)—even though it seems to ignore the multiple-comparisons and post-selection issues. Loosely speaking, when the winner is so clearly different from the runner-up and other candidates, then removing the winner from consideration has negligible statistical effect on our inferences about the rest of the ranking. Admittedly, Hung and Fithian (2016)'s result is unlikely to apply for CV, because we expect many models to fall in the confidence set. When the winner is not significantly different from the runner-up or several others, we cannot ignore the post-selection issues caused by naively using the winner as our baseline in a sequence of tests.

Nonetheless, for CV with confidence, we suggest the following simple "HF" (Hung-Fithian) procedure inspired by their result. For each non-winning model, compute the vector of individual test-case differences in loss between the winner and the other model. Using these difference vectors, create Bonferroni-corrected

2-sided Gaussian confidence intervals for the differences in $\widehat{MSE}$ between the winner and every other model. Let our confidence set contain the winner and any model whose CI for the difference includes 0. We also propose another variant, "HFnaive", in which the CIs are uncorrected and 1-sided (as one might naively do when ignoring the multiple-comparisons and post-selection problems).

The 1SE rule Besides the above ranking-with-confidence approaches, we also consider variants of the "one standard error rule" (1SE), introduced in Breiman et al. (1984) as a heuristic regularization method for model selection. In the original rule, we choose the smallest (or most-regularized) model whose test MSE falls no more than one winner's-SE away from the winner's MSE: $\widehat{MSE}(J) \leq \widehat{MSE}(\hat{J}_{winner}) + \widehat{SE}(\hat{J}_{winner})$. We can turn this into a confidence set by choosing all such models, not just the smallest.

However, this confidence set does not allow us to tune α, and it completely ignores the highly-positive correlations we expect to see between models. Hence, as another approximation to the HF rule above, we propose a method "ZSEdiff" which requires barely any more computation than the 1SE rule does. The SE of a difference is $\sqrt{SE_1^2 + SE_2^2 - 2Cov_{1,2}}$. If we make a drastic simplifying assumption in the spirit of the 1SE rule, and we assume that all models' SEs and pairwise correlations are the same, then we can approximate the SE of any difference between models as $SE_1 \cdot \sqrt{2(1 - Corr_{1,2})}$. In this case, we estimate $\widehat{SE}(\text{diff}) = \widehat{SE}(\hat{J}_{winner}) \cdot \sqrt{2(1 - \widehat{Corr}(\hat{J}_{winner}, \hat{J}_{neighbor}))}$, using the estimated correlation of the winning model's test-case losses with those of the next-smallest or next-largest model—whichever gives the larger $\widehat{SE}(\text{diff})$. We use this $\widehat{SE}(\text{diff})$ to form Bonferroni-corrected 2-sided Gaussian CIs for the differences between the winner and each model, and our confidence set includes any models whose CI for the difference includes 0.

Estimating $\widehat{SE}$ Each of these five proposed methods—KWW, HF, HFnaive, 1SE, and ZSEdiff—requires standard error estimates. The naive SE estimator for sample-splitting takes the standard deviation of individual test-case losses, then divides by $\sqrt{n_{test}}$. Likewise, for V-fold CV, the naive SE estimator takes the standard deviation of average MSEs across folds, then divides by $\sqrt{V}$.

These naive SE estimators are clearly not appropriate. They assume that the test losses are independent across cases and/or across folds. However, in sample-splitting, every test case is evaluated on model estimates from the same training set. A more reasonable covariance structure for ℓ_{im} would be Compound Symmetry, a.k.a. an Exchangeable covariance matrix: σ_d^2 for every diagonal entry, σ_o for every off-diagonal entry. Likewise, a better covariance structure for V-fold CV would consist of such Exchangeable blocks within each fold, and a third constant covariance σ_b between cases in different folds. Nadeau and Bengio (2000) and Bengio and Grandvalet (2004) study estimation of the variance of MCCV and V-fold generalization error estimates, respectively. They propose alternate estimators, but show that this is a challenging problem with no unbiased estimators.

As a starting point, our initial simulations reported below use sample-splitting and rely on the naive SE estimator. Future work will explore how sensitive our methods are to the choice of SE estimator.

Simulations For our initial simulations, we chose to use sample-splitting, on a fixed model path, with orthogonal design. This will make analytical exploration of any interesting empirical findings more tractable.

Data are generated from a linear model with orthogonal Gaussian design and iid Gaussian noise of variance $\sigma^2 = 1$. The regression coefficients are k-sparse with $\beta_{min} = 0.2$. We compute mean test MSEs from a single 80:20 split, on a fixed correct path (the first k models are adding correct variables). Any SEs for the mean test MSE (or the mean difference in squared errors) are computed naively, as the standard deviation divided by $\sqrt{n_{test}}$. We simulate a range of sample sizes $n = 50, 250, 1250$, dimensions $p = 11, 50$, and true model sizes $k = 3, 10$. We also use two confidence levels, $1 - \alpha = 0.95$ and 0.90, in order to assess whether each method's properties are sensitive to the target level.

Besides confidence set coverage and size, we are also interested in model-selection by the "bet on sparsity" approach of choosing the smallest model in the confidence set. At each simulation setting, we generate 1000 datasets and run all 5 methods. Figure 7.1 reports each set's coverage (how often does the set contain the true model?), probability of correct selection (how often is the smallest model in the set actually the true model?), and size (how many models are in the set?).

Except at the smallest n, 1SE is actually quite conservative. This seemed surprising at first—the naive SEs tend to be too small, and only one SE should not be enough. However, because there are such high correlations between models, the appropriate margin of error for a difference between models is actually much smaller than 1 naive SE, leading to conservative coverage for the 1SE rule. KWW's coverage is also very conservative, as expected, but so is HF's. The 2-sided Bonferroni correction appears to be excessive compared to HFnaive. 1SE, KWW, and HF all have negligible probabilities of correct selection, and their sets are much too large to be useful.

HFnaive and ZSEdiff perform surprisingly well. Except at the smallest n, their coverage is close to nominal, and it does change as we adjust the target confidence level. Their probability of correct selection is not high, but it is far better than the other methods. They also tend to have the smallest confidence sets.

Discussion We have illustrated the use of several methods, inspired by approaches to ranking with confidence, for approximating Lei's CVC at much lower computational expense. Two of these methods—HFnaive and ZSEdiff—appear to have promising properties in terms of confidence set coverage, size, and model selection, at least in this simple setting with orthogonal Gaussian design and a fixed model path. A better analytical understanding appears to be worth pursuing, along with further simulation under more diverse data conditions and changes to the methodology. Next steps will also include direct comparison to CVC in terms of coverage and computation time, as well as the use of better SE estimators.

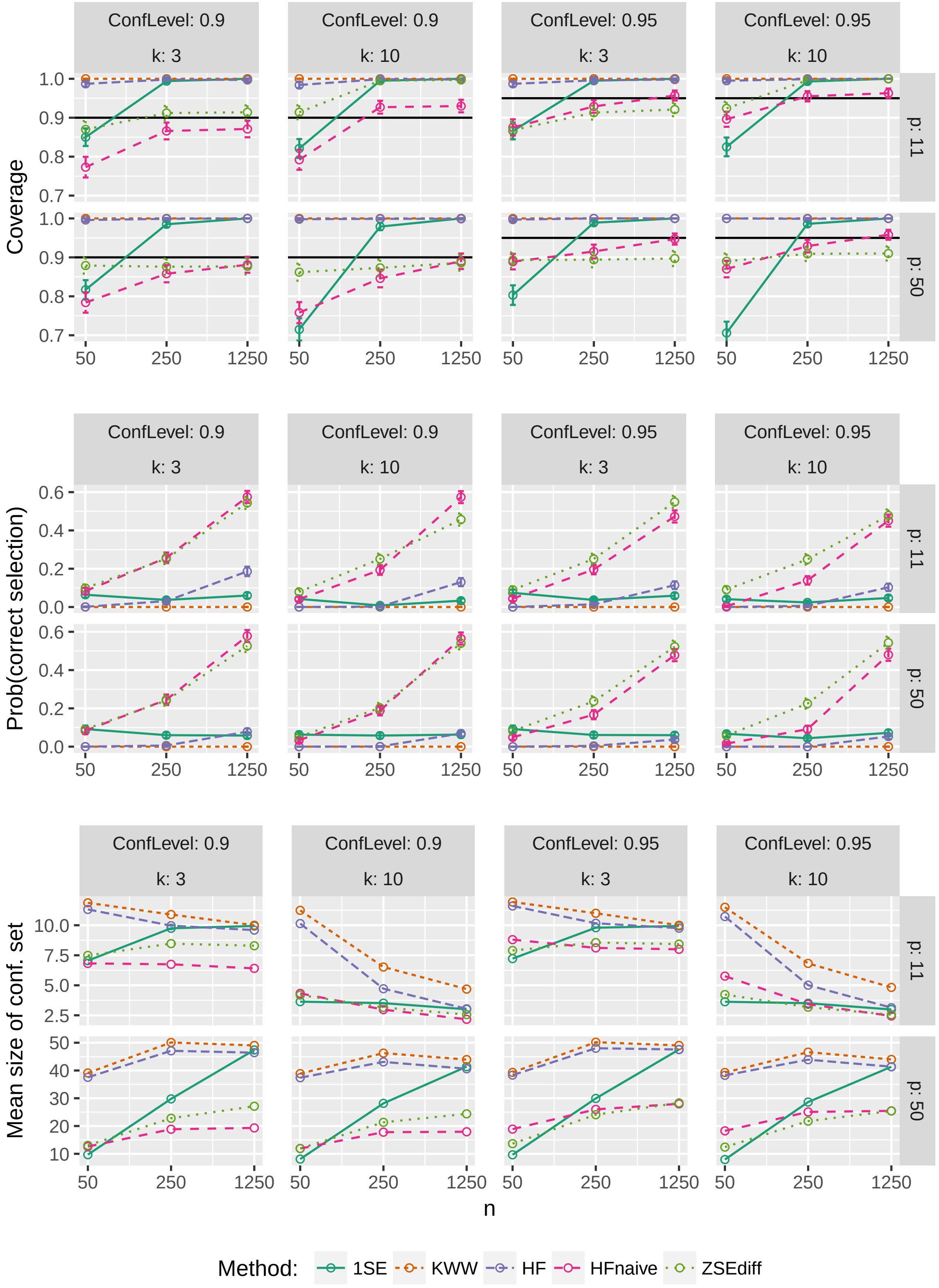

Figure 7.1: Coverage, probability of correct model selection, and average confidence set size for the 5 methods proposed. Based on 1000 simulations at each data point. Error bars show $\pm 2 \cdot SE$.

Chapter 8

Proofs and Lemmas

8.1 Proofs

8.1.1 Proof of Proposition 3.3

For clarity, the derivations below assume a single spurious predictor, so that $p - k = 1$. For the case where $p - k > 1$, apply the same derivations to each spurious variable separately, but using μ and γ computed on the whole dataset (not just with that one spurious variable). We see that under our sufficient conditions, which depend on the spurious predictors only through μ and γ, each step of FS must choose a true variable before any spurious variable. So the proof continues to hold when $p - k > 1$.

First step: $t = 0$

Begin with the first step, $t = 0$. Let $\rho_{j,\ell}$ denote the correlation between columns x_j and x_ℓ for $j \neq \ell \in 1, \ldots, k+1$. This is equivalent to the coherence or inner product $\langle x_j, x_\ell \rangle$ since each column has zero mean and unit norm. Also let $\rho_{j,\epsilon} = \langle x_j, \mathrm{E}/\|\mathrm{E}\| \rangle$, the coherence between x_j and E (not exactly a correlation because we do not assume that E has zero sample mean.)

A sufficient correct decision would be for FS to choose x_1 over the spurious x_{k+1}, which happens if $|\langle x_1, y \rangle| > |\langle x_{k+1}, y \rangle|$, where

$$\langle x_1, y \rangle = \beta_1 + \sum_{j=2}^{k} \beta_j \rho_{1,j} + \|\mathrm{E}\| \rho_{1,\epsilon} \quad \text{and} \quad \langle x_{k+1}, y \rangle = \sum_{j=1}^{k} \beta_j \rho_{k+1,j} + \|\mathrm{E}\| \rho_{k+1,\epsilon} \, .$$

A sufficient condition would be

$$|\beta_1| > |\beta_1| \cdot |\rho_{k+1,1}| + \sum_{j=2}^{k} |\beta_j| (|\rho_{1,j}| + |\rho_{k+1,j}|) + \|\mathrm{E}\| (|\rho_{1,\epsilon}| + |\rho_{k+1,\epsilon}|)$$

which is implied by

$$|\beta_1|(1-\mu-2(k-1)\mu) > \|\mathrm{E}\|2\gamma$$

itself implied by

$$|\beta_1|/\|\mathrm{E}\| > \frac{2\gamma}{1-(2k-1)\mu}\,.$$

Second step: $t = 1$

Assuming we chose x_1 correctly before, now we will correctly choose x_2 over x_{k+1} if

$$\frac{|\langle x_2, R(y|x_1)\rangle|}{\|R(x_2|x_1)\|} > \frac{|\langle x_{k+1}, R(y|x_1)\rangle|}{\|R(x_{k+1}|x_1)\|}\,.$$

We have

$$R(y|x_1) = \sum_{j=2}^{k} \beta_j(x_j - \rho_{1,j}x_1) + \|\mathrm{E}\|(\mathrm{E} - \rho_{1,\epsilon}x_1)$$

$$\langle x_2, R(y|x_1)\rangle = \sum_{j=2}^{k} \beta_j(\rho_{2,j} - \rho_{1,j}\rho_{1,2}) + \|\mathrm{E}\|(\rho_{2,\epsilon} - \rho_{1,\epsilon}\rho_{1,2})$$

$$\|R(x_2|x_1)\| = \sqrt{1-\rho_{1,2}^2}$$

and analogously for the x_{k+1} terms. So we choose correctly if

$$\frac{\sum_{j=2}^{k} \beta_j(\rho_{2,j} - \rho_{1,j}\rho_{1,2}) + \|\mathrm{E}\|(\rho_{2,\epsilon} - \rho_{1,\epsilon}\rho_{1,2})}{\sqrt{1-\rho_{1,2}^2}} > \frac{\sum_{j=2}^{k} \beta_j(\rho_{k+1,j} - \rho_{1,j}\rho_{1,k+1}) + \|\mathrm{E}\|(\rho_{k+1,\epsilon} - \rho_{1,\epsilon}\rho_{1,k+1})}{\sqrt{1-\rho_{1,k+1}^2}}\,.$$

This is implied by

$$\frac{|\beta_2|}{\|\mathrm{E}\|} > \frac{\frac{|\rho_{2,\epsilon}-\rho_{1,\epsilon}\rho_{1,2}|}{\sqrt{1-\rho_{1,2}^2}} + \frac{|\rho_{k+1,\epsilon}-\rho_{1,\epsilon}\rho_{1,k+1}|}{\sqrt{1-\rho_{1,k+1}^2}}}{\sqrt{1-\rho_{1,2}^2} - \sum_{j=3}^{k} \frac{|\rho_{2,j}-\rho_{1,j}\rho_{1,2}|}{\sqrt{1-\rho_{1,2}^2}} - \sum_{j=2}^{k} \frac{|\rho_{k+1,j}-\rho_{1,j}\rho_{1,k+1}|}{\sqrt{1-\rho_{1,k+1}^2}}}\,.$$

We can maximize the right-hand side by plugging in μ for ρ in the square-root terms to get a simpler sufficient condition:

$$\frac{|\beta_2|}{\|\mathrm{E}\|} > \frac{|\rho_{2,\epsilon} - \rho_{1,\epsilon}\rho_{1,2}| + |\rho_{k+1,\epsilon} - \rho_{1,\epsilon}\rho_{1,k+1}|}{1-\mu^2 - |\rho_{k+1,2} - \rho_{1,2}\rho_{1,k+1}| - \sum_{j=3}^{k}(|\rho_{2,j} - \rho_{1,j}\rho_{1,2}| + |\rho_{k+1,j} - \rho_{1,j}\rho_{1,k+1}|)}\,. \tag{8.1}$$

The RHS numerator of (8.1) has the bound $Num \leq 2\gamma(1+\mu)$, which is increasing with μ. In the RHS denominator, $|\rho_a - \rho_b\rho_c| \leq |\mu + \mu^2| = \mu(1+\mu)$ gives a bound decreasing in μ:

$$Den \geq 1-\mu^2 - (2(k-2)+1)\mu(1+\mu) = 1-\mu^2 - (2k-3)\mu(1+\mu) = (1+\mu)(1-(2k-2)\mu)$$

which is strictly positive if $\mu < (2k-1)^{-1}$.

Combining these numerator and denominator bounds gives a sufficient condition for (8.1):

$$\frac{|\beta_2|}{\|\mathrm{E}\|} > \frac{2\gamma(1+\mu)}{(1+\mu)(1-(2k-2)\mu)} = \frac{2\gamma}{1-(2k-2)\mu}\,.$$

Later steps: $t > 1$

We introduce new notation for the remaining steps. Imagine adding a rescaled E as the final column of the design matrix: $x_{k+2} = \mathrm{E}/\|\mathrm{E}\|$. FS still cannot choose it as a predictor, but this will help us track it in our derivations.

Assume that so far FS has correctly added the predictor set $J_t = \{1,\ldots,t\}$ to the model. Define the residuals at step t as $y_t = R(y|\mathbf{x}_{J_t})$ and $x_{j,t} = R(x_j|\mathbf{x}_{J_t})$ for $j > t$. Decompose the response residual into the sum of a signal residual and noise residual: $y_t = S_t + N_t$, where $S_t = R(\beta_1 x_1 + \ldots + \beta_k x_k|\mathbf{x}_{J_t})$ and $N_t = R(\mathrm{E}|\mathbf{x}_{J_t}) = \|\mathrm{E}\| \cdot R(x_{k+2}|\mathbf{x}_{J_t})$.

Now conduct a QR decomposition of this augmented design matrix: $(x_1,\ldots,x_{k+2}) = (Z_1,\ldots,Z_{k+2})A$, where the Z_i columns are orthonormal and A is an upper triangular matrix with positive diagonal entries. Thus, the coherence matrix $C = \mathbf{x}^T\mathbf{x}$ can also be written as $C = A^TA$, i.e. the Cholesky decomposition of C. Let A_J be the principal submatrix using index set J and let $a_{i,j}$ be the entry in A's row i, column j.

Notice that:

$$\begin{aligned}
S_t &= (Z_{t+1},\ldots,Z_k)A_{t+1:k}\beta_{t+1:k} \\
x_{t+1,t} &= Z_{t+1}a_{t+1,t+1} \\
\|x_{t+1,t}\| &= a_{t+1,t+1} \\
x_{k+1,t} &= [(Z_{t+1},\ldots,Z_{k+1})A_{t+1:k+1}]_{k+1} \\
N_t &= \|\mathrm{E}\| \cdot [(Z_{t+1},\ldots,Z_{k+2})A_{t+1:k+2}]_{k+2}\,.
\end{aligned}$$

For FS to correctly choose $t+1$ next instead of the spurious $k+1$, we need

$$\frac{|\langle S_t + N_t, x_{t+1,t}\rangle|}{\|x_{t+1,t}\|} > \frac{|\langle S_t + N_t, x_{k+1,t}\rangle|}{\|x_{k+1,t}\|}$$

for which a sufficient condition is

$$\frac{|\langle S_t, x_{t+1,t}\rangle|}{\|x_{t+1,t}\|} - \frac{|\langle N_t, x_{t+1,t}\rangle|}{\|x_{t+1,t}\|} > \frac{|\langle S_t, x_{k+1,t}\rangle|}{\|x_{k+1,t}\|} + \frac{|\langle N_t, x_{k+1,t}\rangle|}{\|x_{k+1,t}\|}\,.$$

We can rewrite each term as follows:

$$\begin{aligned}
\frac{\langle S_t, x_{t+1,t}\rangle}{\|x_{t+1,t}\|} &= \frac{\sum_{j=t+1}^{k} a_{t+1,j}a_{t+1,t+1}\beta_j}{a_{t+1,t+1}} = \sum_{j=t+1}^{k} a_{t+1,j}\beta_j \\
\frac{\langle S_t, x_{k+1,t}\rangle}{\|x_{k+1,t}\|} &= \frac{\sum_{j=t+1}^{k}\left(\sum_{l=t+1}^{j} a_{l,j}a_{l,k+1}\right)\beta_j}{\|x_{k+1,t}\|} = \sum_{j=t+1}^{k} \frac{\langle x_{k+1,t}, x_{j,t}\rangle}{\|x_{k+1,t}\|}\beta_j \\
\frac{\langle N_t, x_{t+1,t}\rangle}{\|x_{t+1,t}\|} &= \frac{\|\mathrm{E}\|}{a_{t+1,t+1}} \cdot \langle a_{t+1,t+1}Z_{t+1}, [Z_{t+1:k+2}A_{t+1:k+2}]_{k+2}\rangle = \|\mathrm{E}\| a_{t+1,k+2} \\
\frac{\langle N_t, x_{k+1,t}\rangle}{\|x_{k+1,t}\|} &= \frac{\|\mathrm{E}\|}{\|x_{k+1,t}\|} \cdot \langle [Z_{t+1:k+2}A_{t+1:k+2}]_{k+2}, [Z_{t+1:k+1}A_{t+1:k+1}]_{k+1}\rangle = \|\mathrm{E}\| \cdot \frac{\langle x_{k+1,t}, x_{k+2,t}\rangle}{\|x_{k+1,t}\|}\,.
\end{aligned}$$

This gives the sufficient condition

$$|\beta_{t+1}| \cdot |a_{t+1,t+1}| > |\beta_{t+1}| \left(\sum_{j=t+2}^{k} |a_{t+1,j}| + \sum_{j=t+1}^{k} \frac{|\langle x_{k+1,t}, x_{j,t}\rangle|}{\|x_{k+1,t}\|} \right) + \|\mathrm{E}\| \left(|a_{t+1,k+2}| + \frac{|\langle x_{k+1,t}, x_{k+2,t}\rangle|}{\|x_{k+1,t}\|} \right).$$

Now we use Lemma 8.1, on subsets and rearrangements of C, to lower-bound $a_{t+1,t+1} \geq \sqrt{\frac{1-t\mu}{1-(t-1)\mu}}$. We also upper-bound $a_{t+1,j}$ and $\frac{\langle x_{k+1,t}, x_{j,t}\rangle}{\|x_{k+1,t}\|}$ by $\frac{\mu}{1-(t+1)\mu}$ when $j \in \{t+2, \ldots, k\}$. And we upper-bound these same two terms by $\frac{\gamma}{1-t\mu-(t+1)\gamma^2}$ when $j = k+2$.

(The reason Lemma 8.1 applies to $\frac{\langle x_{k+1,t}, x_{j,t}\rangle}{\|x_{k+1,t}\|}$ is that this is the $(t+1, t+2)$ term in the Cholesky decomposition of $(x_{1:t}, x_{k+1}, x_j)^T(x_{1:t}, x_{k+1}, x_j)$.)

Plugging in these bounds, we get the sufficient condition

$$|\beta_{t+1}| \sqrt{\frac{1-t\mu}{1-(t-1)\mu}} > (2k-2t-1)|\beta_{t+1}| \frac{\mu}{1-(t+1)\mu} + 2\|\mathrm{E}\| \frac{\gamma}{1-t\mu-(t+1)\gamma^2}\,.$$

Therefore, FS will make a correct choice at each $t \geq 2$ if the signal-to-noise ratio is at least

$$\frac{|\beta_{t+1}|}{\|\mathrm{E}\|} > \frac{\frac{2\gamma}{1-t\mu-(t+1)\gamma^2}}{\sqrt{\frac{1-t\mu}{1-(t-1)\mu}} - \frac{(2k-2t-1)\mu}{1-(t+1)\mu}}\,. \tag{8.2}$$

8.1.2 Proof of Corollary 3.4

For $t = 1$, note that

$$\frac{2\gamma}{1-(2k-2)\mu} < \frac{2\gamma}{1-\frac{2k-2}{2k-1}} = 2\gamma(2k-1) < 4k\gamma$$

so that $4k\gamma$ is a sufficient lower bound on the signal-to-noise ratio $\frac{|\beta_2|}{\|\mathrm{E}\|}$. This also holds for $t = 0$ since we assume $|\beta_1| \geq |\beta_2|$.

For $t > 1$, the RHS of equation (8.2) has a numerator increasing in t. So we can upper-bound the RHS by plugging in the largest relevant value: $t = k - 1$.

Meanwhile, Lemma 8.2 shows the RHS denominator is also increasing in t for $0 < \mu < (2k-1)^{-1}$, $t \geq 2$, $k \geq 3$. So we can upper-bound the RHS by plugging in the smallest relevant value: $t = 2$.

This gives a sufficient condition in γ, μ, k that holds across all $t \in 2, \ldots, k-1$:

$$\min_{j \in 1,\ldots,k} \frac{|\beta_j|}{\|\mathrm{E}\|} > \frac{2\gamma}{(1-(k-1)\mu - k\gamma^2)\left(\sqrt{\frac{1-2\mu}{1-\mu}} - \frac{(2k-5)\mu}{1-3\mu}\right)} .$$

Note that, in the denominator,

$$1 - (k-1)\mu - k\gamma^2 > 1 - \frac{k-1}{2k-1} - k(2k-1)^{-2} = k\frac{2k-2}{(2k-1)^2} .$$

Its inverse is approximately 2, bounded above by $25/12 \approx 2.1$ when $k = 3$ (and below by 2 as $k \to \infty$). So our bound becomes

$$\min_{j \in 1,\ldots,k} \frac{|\beta_j|}{\|\mathrm{E}\|} \geq \frac{4.2\gamma}{\sqrt{\frac{1-2\mu}{1-\mu}} - \frac{(2k-5)\mu}{1-3\mu}} .$$

Now the denominator is decreasing in μ, so we can plug in the worst case $\mu = (2k-1)^{-1}$:

$$\min_{j \in 1,\ldots,k} \frac{|\beta_j|}{\|\mathrm{E}\|} \geq \frac{4.2\gamma}{\sqrt{\frac{2k-3}{2k-2}} - \frac{2k-5}{2k-4}} = \frac{4.2\gamma}{\sqrt{1 - \frac{1}{2k-2}} - \left(1 - \frac{1}{2k-4}\right)} .$$

Using a Maclaurin series for $(1-x)^{1/2} \approx 1 - \frac{x}{2}$ (in fact $\sqrt{1-x} \leq 1 - \frac{x}{2}$ for $0 < x < 1$ since $\sqrt{1-x}$ is monotonically decreasing) and evaluating it at $x = (2k-2)^{-1}$, we get

$$\min_{j \in 1,\ldots,k} \frac{|\beta_j|}{\|\mathrm{E}\|} \geq \frac{4.2\gamma}{\frac{1}{2k-4} - \frac{1}{2(2k-2)}} = \frac{8.4\gamma}{\frac{1}{k-2} - \frac{1}{2k-2}} \approx \frac{8.4\gamma}{\frac{1}{k}(1-\frac{1}{2})} = 16.8k\gamma .$$

8.1.3 Proof of Corollary 3.7

Repeat the proof of Proposition 3.3, but use the s-sparse results from Cases 1 and 2 of Lemma 8.1. We get

$$\sum_{j=t+2}^{k} |a_{t+1,j}| \leq \mathrm{rowsum}_j(|A - I|) \leq \frac{s\mu}{1-s\mu}$$

where we have k playing the role of p. Since we already assume that $s < k$, we can just use s instead of $\min\{s, k-1\}$. The same argument works for the other term:

$$\sum_{j=t+1}^{k} \frac{|\langle x_{k+1,t}, x_{j,t}\rangle|}{\|x_{k+1,t}\|} \leq \frac{s\mu}{1-s\mu} .$$

All together, our sufficient condition becomes

$$|\beta_{t+1}|\sqrt{\frac{1-\min\{s,t\}\mu}{1-(\min\{s,t\}-1)\mu}} \ > \ |\beta_{t+1}|\frac{2s\mu}{1-s\mu}+2\|\mathrm{E}\|\frac{\gamma}{1-\min\{s,t\}\mu-(t+1)\gamma^2}$$

which avoids the use of k, as desired. Note that for fixed feasible s and μ, the LHS is smallest when $t \geq s$, and the RHS 2nd term is also largest when $t \geq s$. So a sufficient condition would be

$$|\beta_{t+1}|\sqrt{\frac{1-s\mu}{1-(s-1)\mu}} \ > \ |\beta_{t+1}|\frac{2s\mu}{1-s\mu}+2\|\mathrm{E}\|\frac{\gamma}{1-s\mu-(t+1)\gamma^2}$$

or equivalently

$$\frac{|\beta_{t+1}|}{\|\mathrm{E}\|} > \frac{2\gamma}{(1-s\mu-(t+1)\gamma^2)\left(\sqrt{\frac{1-s\mu}{1-(s-1)\mu}}-\frac{2s\mu}{1-s\mu}\right)}$$

as long as $\sqrt{\frac{1-s\mu}{1-(s-1)\mu}}-\frac{2s\mu}{1-s\mu}>0$. The condition $\mu<(3.4s)^{-1}$ is sufficient for this to hold for any $1 \leq s < p$: Plug in $\mu=(3.4s)^{-1}$ to see that even if $s=1$,

$$\sqrt{\frac{1-s\mu}{1-(s-1)\mu}}-\frac{2s\mu}{1-s\mu}=\sqrt{\frac{2.4s}{2.4s+1}}-\frac{2}{2.4} \geq \sqrt{\frac{2.4}{3.4}}-\frac{2}{2.4} \approx 0.0068 > 0\,.$$

(The 3.4 approximates the solution to $(x-1)^3-4x=0$, whose exact form is not simple.)

Assuming $\mu<(3.4s)^{-1}$, and defining $q(s) \equiv 2\cdot\left(\sqrt{\frac{2.4s}{2.4s+1}}-\frac{2}{2.4}\right)^{-1}$, we can simplify to the step-by-step sufficient condition

$$\frac{|\beta_{t+1}|}{\|\mathrm{E}\|} > \frac{2\gamma}{\left(\frac{2.4}{3.4}-(t+1)\gamma^2\right)\left(\sqrt{\frac{2.4s}{2.4s+1}}-\frac{2}{2.4}\right)}=\frac{\gamma\cdot q(s)}{\frac{2.4}{3.4}-(t+1)\gamma^2}$$

or further to the across-all-steps condition

$$\frac{|\beta_{min}|}{\|\mathrm{E}\|} > \frac{\gamma\cdot q(s)}{\frac{2.4}{3.4}-k\gamma^2}=\frac{\gamma\cdot q(s)}{\frac{12}{17}-k\gamma^2}$$

as long as $\mu<(3.4s)^{-1}\approx\frac{0.29}{s}$ and $\gamma<\sqrt{\frac{12}{17k}}\approx\frac{0.84}{\sqrt{k}}$.

$q(s)$ is greatest for small s, as $q(1)\approx 293$, but asymptotes towards $q(s)\approx 12$ as $s\to\infty$.

8.1.4 Proof of Proposition 3.8

Assume each element of E has mean 0 and variance σ^2/n and is i.i.d. from some sub-Gaussian distribution. For $j=1,\ldots,p$, let $W_j=\langle x_j,\mathrm{E}\rangle$. Then $\mathbb{E}(W_j)=0$ and $\mathbb{V}(W_j)=\frac{\sigma^2}{n}\cdot\|x_j\|_2^2=\frac{\sigma^2}{n}$, since the columns of $\mathbf{x}$ have unit norm.

Thus, each $\frac{\sqrt{n}}{\sigma} \cdot W_j$ is a linear combination of sub-Gaussians and therefore sub-Gaussian itself, with constant (unit) variance. By the union bound and the sub-Gaussian tail inequality from Lemma 8.3, there exist constants $c_1, c_2 > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\max_{j=1,\ldots,p} |\langle x_j, \mathrm{E}\rangle| > \frac{\delta\sigma}{\sqrt{n}}\right) &= \mathbb{P}\left(\max_{j=1,\ldots,p} \frac{\sqrt{n}}{\sigma}|W_j| > \delta\right) \\ &\leq p \cdot \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}|W_j| > \delta\right) \\ &\leq pc_1 e^{-c_2\delta^2} .\end{aligned}$$

If we choose $\eta > 0$ and $\delta = \sqrt{\frac{(1+\eta)\log(p)}{c_2}}$, then

$$\mathbb{P}\left(\max_{j=1,\ldots,p} |\langle x_j, \mathrm{E}\rangle| > \sigma\sqrt{\frac{\log(p)}{n}} \cdot \sqrt{\frac{1+\eta}{c_2}}\right) \leq \frac{c_1}{p^\eta} .$$

So, for large p, we have $\hat{\gamma}\|\mathrm{E}\| = \max_{j=1,\ldots,p} |\langle x_j, \mathrm{E}\rangle| = O\left(\sigma\sqrt{\log(p)/n}\right)$ with high probability of at least $1 - c_1 p^{-\eta}$.

8.1.5 Proof of Proposition 3.10

Let $S = n^{-1}(\mathbf{X} - \overline{\mathbf{X}})^T(\mathbf{X} - \overline{\mathbf{X}})$ be the sample covariance matrix of $\mathbf{X}$, and let C be the corresponding sample correlation matrix.

Lemma 8.4 shows that $\|S - \Sigma\|_{\infty,\infty} = O(\sqrt{\log(p)/n})$ with high probability. Lemma 8.5 extends this to $\|C - \Sigma\|_{\infty,\infty}$, as well as to versions of these matrices augmented with an extra row & column for the (unstandardized) noise ϵ.

First, for a given observed sample, $\hat{\mu}$ is the highest coherence in the dataset, achieved by some pair of variables. Let $\mu_\bullet$ denote the entry of Σ corresponding to this same pair of variables. Obviously $\mu_\bullet \leq \mu$. By Lemma 8.5, with high probability, $\|C - \Sigma\|_{\infty,\infty}$ has an upper bound $b_n = O\left(\sqrt{\frac{\log p}{n}}\right)$ which shrinks as n grows. Thus for a large enough n, we have $|\hat{\mu} - \mu_\bullet| < b_n < (2k-1)^{-1} - \mu$ and so $\hat{\mu} < (2k-1)^{-1}$, with high probability.

Second, Lemma 8.5 shows that $\hat{\gamma}\|\epsilon\|/\sqrt{n} = O(\sigma\sqrt{\log(p)/n})$ with high probability.

Each of these "high probabilities" has the form $1 - c_i p^{-\eta}$ for some $c_1, c_2 > 0$ and our choice of $\eta > 0$. By the union bound, both events occur at once with probability at least $1 - (c_1 + c_2)p^{-\eta}$.

8.1.6 Proof of Proposition 4.5

Here we are primarily working with submodels of the true model $J_* = \{1, \ldots, k\}$. In this section, we will use J_h to denote any one of these 2^k possible submodels, and we index these models using $h \in 1, \ldots, 2^k$.

$\mathbf{X}_c$ and $\mathbf{X}_v$ are respectively the full construction and validation sets. $\mathbf{X}_{c,J_h}$ contains just the training observations for columns in model J_h.

Proof sketch

Under the conditions of Theorem 3.1, we claim that the probability of underfit goes to 0 as n, k grow and β^2_{min} shrinks, as long as $\beta^2_{min} \geq g(\beta_{max}, k, n_c, \sigma) \equiv c \cdot \max\left\{\beta^2_{max}\sqrt{\frac{k^2\log(k)}{n_c}}, |\beta_{max}|\sigma k\sqrt{\frac{\log(k)}{n_v}}\right\}$ for some $c > 0$.

For each model J_h (for $h \in 1, \ldots, 2^k$), we decompose its CV estimate of MSE into signal b_{J_h} and noise: $\widehat{MSE}(J_h) = b_{J_h} + \nu_{J_h} + r$, where r does not depend on h. Then we show that:

- By the conditions of Theorem 3.1 and proof of Proposition 3.3, with probability at least $1 - \gamma_1(p) \to 1$, FS will choose the next predictor $\hat{j}$ such that:
    - $\hat{j} \in J_*$, i.e. it is a correct variable;
    - the training-estimate of the risk improves over model J_h, that is, $\widehat{Risk}_c(\hat{\beta}_{J_h \cup \hat{j}}) < \widehat{Risk}_c(\hat{\beta}_{J_h})$; and
    - the difference in signals is at least $\Delta(\beta_{min})$, that is, $b_{J_h \cup \hat{j}} < b_{J_h} - \Delta(\beta_{min})$, uniformly over all J_h strictly smaller than J_*.
- With probability at least $1 - \gamma_2(k) \to 1$, the maximum noise term magnitude $\max_h |\nu_{J_h}|$ is less than $\frac{1}{2}\Delta(\beta_{min})$, as long as $\beta^2_{min} > g(\beta_{max}, k, n_c, \sigma)$.

Therefore, $\beta^2_{min} > g(\beta_{max}, k, n_c, \sigma)$ implies that the testing-estimate of the risk also improves and therefore FS does not stop at model J_h, uniformly over h and with high probability:

$$\mathbb{P}\left(\widehat{Risk}_v(\hat{\beta}_{J_h \cup \hat{j}}) \leq b_{J_h \cup \hat{j}} + \max_h |\nu_{J_h}| + r < b_{J_h} - \max_h |\nu_{J_h}| + r \leq \widehat{Risk}_v(\hat{\beta}_{J_h})\right) \to 1$$

or

$$\mathbb{P}\left(\min_h\left(\widehat{MSE}(J_h) - \widehat{MSE}(J_h \cup \hat{j})\right) \leq 0\right) \leq \mathbb{P}\left(\Delta(\beta_{min}) < 2\max_h |\nu_{J_h}|\right) \to 0$$

as $n \to \infty$. Specifically,

$$\begin{aligned}\mathbb{P}(\text{CV chooses underfit model}) &\leq \mathbb{P}(\text{FS chooses incorrect path}) \\ &\quad + \gamma_2(k)\mathbb{P}(\text{FS chooses correct path}) \\ &\leq \gamma_1(p) + \gamma_2(k)(1 - \gamma_1(p)) \\ &\to 0\,.\end{aligned}$$

Decompose $\widehat{MSE}(J_h)$

$$\begin{aligned}\widehat{MSE}(J_h) =&(\beta - \hat{\beta}_{J_h})^T \frac{\mathbf{X}_v^T\mathbf{X}_v}{n_v}(\beta - \hat{\beta}_{J_h}) + \frac{\epsilon_v^T\epsilon_v}{n_v} + 2\frac{\epsilon_v^T\mathbf{X}_v}{n_v}(\beta - \hat{\beta}_{J_h}) \\ =&(\beta - \hat{\beta}_{J_h})^T \frac{\mathbf{X}_c^T\mathbf{X}_c}{n_c}(\beta - \hat{\beta}_{J_h}) + \frac{\epsilon_v^T\epsilon_v}{n_v} \\ &+ \left[(\beta - \hat{\beta}_{J_h})^T \left(\frac{\mathbf{X}_v^T\mathbf{X}_v}{n_v} - \frac{\mathbf{X}_c^T\mathbf{X}_c}{n_c}\right)(\beta - \hat{\beta}_{J_h}) + 2\frac{\epsilon_v^T\mathbf{X}_v}{n_v}(\beta - \hat{\beta}_{J_h})\right] .\end{aligned}$$

Let $P_h = \mathbf{X}_{c,J_h}(\mathbf{X}_{c,J_h}^T\mathbf{X}_{c,J_h})^{-1}\mathbf{X}_{c,J_h}^T$ be the construction-set projection onto the columns in model J_h, and

$$\begin{aligned}(\beta - \hat{\beta}_{J_h})^T \frac{\mathbf{X}_c^T\mathbf{X}_c}{n_c}(\beta - \hat{\beta}_{J_h}) &= n_c^{-1}\|\mathbf{X}_c\beta - P_h(\mathbf{X}_c\beta + \epsilon_c)\|^2 \\ &= n_c^{-1}\left(\|(I - P_h)\mathbf{X}_c\beta\|^2 + \|P_h\epsilon_c\|^2\right) .\end{aligned}$$

Therefore, $\widehat{MSE}(J_h) = b_{J_h} + \nu_{J_h} + r$, where $b_{J_h} = n_c^{-1}\|(I - P_h)\mathbf{X}_c\beta\|^2$; $r = \frac{\epsilon_v^T\epsilon_v}{n_v}$ which cancels out of every comparison $\widehat{MSE}(J_h) - \widehat{MSE}(J_{h'})$; and

$$\nu_{J_h} = (\beta - \hat{\beta}_{J_h})^T \left(\frac{\mathbf{X}_v^T\mathbf{X}_v}{n_v} - \frac{\mathbf{X}_c^T\mathbf{X}_c}{n_c}\right)(\beta - \hat{\beta}_{J_h}) + 2\frac{\epsilon_v^T\mathbf{X}_v}{n_v}(\beta - \hat{\beta}_{J_h}) + n_c^{-1}\|P_h\epsilon_c\|^2 .$$

Lower bound on $b_{J_h} - b_{J_h \cup \hat{j}}$

First, if $J_h = \emptyset$ (i.e. no variable has been chosen yet) and $k = 1$, then $b_\emptyset - b_{\{\hat{j}\}} = n_c^{-1}\|P_{\{\hat{j}\}}\mathbf{X}_c\beta\|^2 = \beta_{min}^2\widehat{\sigma^2}_{X_{\hat{j}}} \geq c \cdot \beta_{min}^2$ with high probability for any choice of $c \in (0, 1)$. Next, assume $k > 1$.

WLOG, reorder columns $1 : k$ so that the variables in model J_h are first and that $\hat{j}$ is next, so $|J_h|+1$ is the index of variable $\hat{j}$. Let A^TA be the Cholesky decomposition of $n_c^{-1}\mathbf{X}_c^T\mathbf{X}_c$ (note that here we do not assume standardized columns of $\mathbf{X}$, which we did in Proposition 3.3 and Lemma 8.1). By the QR decomposition approach in the proof of Proposition 3.3, with high probability the observed sample coherence is below the population bound $\mu < (2k - 1)^{-1}$ and also FS chooses $\hat{j}$ that satisfies

$$\begin{aligned}b_{J_h} - b_{J_h \cup \hat{j}} \geq \left(\sum_{j=|J_h|+1}^{k} \beta_j a_{|J_h|+1,j}\right)^2 &\geq \left(|\beta_{|J_h|+1}a_{|J_h|+1,|J_h|+1}| - \sum_{j=|J_h|+2}^{k} |\beta_j a_{|J_h|+1,j}|\right)^2 \\ &\geq \beta_{|J_h|+1}^2\left(|a_{|J_h|+1,|J_h|+1}| - \sum_{j=|J_h|+2}^{k} |a_{|J_h|+1,j}|\right)^2 .\end{aligned}$$

If $J_h = \emptyset$, then $a_{1,1} = 1$, and by Lemma 8.1, $|a_{1,j}| \leq \frac{\mu}{1-\mu}$ for all other $j = 2, \ldots, k$. Then

$$b_\emptyset - b_{\hat{j}} \geq \beta_{min}^2\left(\frac{1 - k\mu}{1 - \mu}\right)^2$$

This is decreasing in μ, so plug in the upper bound $\mu = (2k-1)^{-1}$:

$$b_{\emptyset} - b_{\hat{j}} \geq \beta_{min}^2 \left(\frac{k-1}{2k-2}\right)^2 = \beta_{min}^2/4$$

so again, $b_{\emptyset} - b_{\hat{j}} \geq c \cdot \beta_{min}^2$ with high probability with $c > 0$.

Otherwise, $J_h \neq \emptyset$. By Lemma 8.1, $|a_{|J_h|+1,|J_h|+1}| \geq \sqrt{\frac{1-|J_h|\mu}{1-(|J_h|-1)\mu}}$ and $|a_{|J_h|+1,j}| \leq \frac{\mu}{1-(|J_h|+1)\mu}$ for all other $j = |J_h|+2, \ldots, k$. So we can lower-bound

$$b_{J_h} - b_{J_h \cup \hat{j}} \geq \beta_{min}^2 \left(\sqrt{\frac{1-|J_h|\mu}{1-(|J_h|-1)\mu}} - \frac{(k-(|J_h|+1))\mu}{1-(|J_h|+1)\mu}\right)^2 .$$

As in Lemma 8.2, the RHS is increasing in $|J_h|$ for $\mu < (2k-1)^{-1}$, so we bound it by plugging in the smallest appropriate $|J_h|$. The remaining case (not yet addressed) is when $|J_h| \geq 1$ and $k \geq 2$, so use $|J_h| = 1$:

$$b_{J_h} - b_{J_h \cup \hat{j}} \geq \beta_{min}^2 \left(\sqrt{1-\mu} - \frac{(k-2)\mu}{1-2\mu}\right)^2 .$$

This is decreasing in μ, so plug in the largest $\mu = (2k-1)^{-1}$:

$$b_{J_h} - b_{J_h \cup \hat{j}} \geq \beta_{min}^2 \left(\sqrt{1 - \frac{1}{2k-1}} - \frac{k-2}{2k-3}\right)^2 .$$

Applying the argument from Corollary 1, for $k \geq 2$ we find that this is strictly decreasing in k and asymptotes towards

$$b_{J_h} - b_{J_h \cup \hat{j}} \geq \beta_{min}^2/4 .$$

Thus, $\min_h b_{J_h} - b_{J_h \cup \hat{j}} = c \cdot \beta_{min}^2 \equiv \Delta(\beta_{min})$. With high probability, FS will always choose $\hat{j}$ whose contribution is at least $c \cdot \beta_{min}^2$, for some $c > 0$.

Upper bound on $|\nu_{J_h}|$ with high probability

Let $\mathbf{X}_{J_h}$ contain all data rows for the columns in model J_h. Note that every such model's covariance matrix satisfies $\|\Sigma_{J_h}\| = O(1)$ and $\|\Sigma_{J_h}^{-1}\| = O(1)$: In the extreme case where Σ_* has constant off-diagonal correlation μ, we have $\Sigma_* = (1-\mu)I_k + \mu \mathbf{1}_k \mathbf{1}_k^T$, whose eigenvalues are $1 + \mu(k-1) < 1 + \frac{k-1}{2k-1} < 1.5$ and $1 - \mu > 1 - \frac{1}{2k-1} > 0.5$, so both $\|\Sigma_*\|, \|\Sigma_*^{-1}\| = O(1)$. These upper and lower bounds also hold for Σ_{J_h} for any sub-model $J_h \subset J_*$.

By Vershynin (2011), Theorem 5.39, let $A_{J_h} = \Sigma_{J_h}^{-1/2} \mathbf{X}_{J_h}$ which is isotropic, and then

$$\mathbb{P}\left(\|A_{J_h}\| < \sqrt{n} + c_1\sqrt{|J_h|} + t\right) \geq 1 - 2\exp(-c_2 t^2) .$$

Choose $t = \sqrt{k/c_2}$, so that for any particular h, $\mathbb{P}\left(\|A_{J_h}\| \geq \sqrt{n}\right) \leq 2\exp(-k)$. Then, union-bounding over all possible models J_h, the probability that at least one norm is "too big" approaches

$$\mathbb{P}\left(\max_h \|A_{J_h}\| \geq \sqrt{n}\right) \leq 2^k \cdot 2\exp(-k) = 2 \cdot (2/e)^{k+1} \to 0$$

as long as k is eventually less than a constant multiple of n (which it must be, since we assume $n^{-1}k^2\log(p) \to 0$).

On the other hand, choosing $t = \sqrt{\log(k)/c_2}$ lets us bound the full matrix with all k columns:

$$\mathbb{P}\left(\|A_*\| \geq \sqrt{n}\right) \leq 2\exp(-\log(k)) = 2k^{-1} \to 0\,.$$

Now let $\hat{\Sigma}_c = \frac{\mathbf{X}_c^T\mathbf{X}_c}{n_c}$. By the above, we have $\mathbb{P}(\|\hat{\Sigma}_c - \Sigma\| \geq c_1\sqrt{\log(k)/n_c}) \leq 2k^{-1}$, and likewise for $\hat{\Sigma}_v$, so

$$\mathbb{P}\left(\left\|\frac{\mathbf{X}_v^T\mathbf{X}_v}{n_v} - \frac{\mathbf{X}_c^T\mathbf{X}_c}{n_c}\right\| \geq c_1\sqrt{\log(k)/n_c}\right) \leq 2k^{-1}\,.$$

Additionally, let $\dot{\beta}_{J_h}$ be the population-version coefficient vector for the best linear approximation using only the variables in model J_h. If we order the columns of $\mathbf{X}$ so that the covariates in J_h come first, then

$$\dot{\beta}_{J_h} = \Sigma^{-1}_{1:|J_h|,\,1:|J_h|}\ \Sigma_{1:|J_h|,\,1:k}\ \beta = \begin{bmatrix} \beta_{1:|J_h|} \\ 0 \end{bmatrix} + \begin{bmatrix} \Sigma^{-1}_{1:|J_h|,\,1:|J_h|}\ \Sigma_{1:|J_h|,\,(|J_h|+1):k}\ \beta_{(|J_h|+1):k} \\ 0 \end{bmatrix}$$

$$\beta - \dot{\beta}_{J_h} = \begin{bmatrix} -\Sigma^{-1}_{1:|J_h|,\,1:|J_h|}\ \Sigma_{1:|J_h|,\,(|J_h|+1):k}\ \beta_{(|J_h|+1):k} \\ \beta_{(|J_h|+1):k} \end{bmatrix}.$$

Then $\dot{\beta}_{J_h} - \hat{\beta}_{J_h}$ has $|J_h| < k$ entries that are $O_p(\sigma|\beta_{max}|\sqrt{\log(k)/n_c})$ each, while $\beta - \dot{\beta}_{J_h}$ has $|J_h| < k$ entries which are at most $O_p(|\beta_{max}|)$ each. Therefore, with probability at least $1 - 2k^{-1}$,

$$\|\beta - \hat{\beta}_{J_h}\|^2 = \|\beta - \dot{\beta}_{J_h} + \dot{\beta}_{J_h} - \hat{\beta}_{J_h}\|^2 = O_p\left(k\beta^2_{max}\left(1 + \sigma\sqrt{\log(k)/n_c}\right)^2\right) = O_p\left(k\beta^2_{max}\right)\,.$$

Then with probability at least $1 - 4k^{-1}$, we have

$$(\beta - \hat{\beta}_{J_h})^T\left(\frac{\mathbf{X}_v^T\mathbf{X}_v}{n_v} - \frac{\mathbf{X}_c^T\mathbf{X}_c}{n_c}\right)(\beta - \hat{\beta}_{J_h}) = \|\beta - \hat{\beta}_{J_h}\|^2 \cdot O\left(\sqrt{\log(k)/n_c}\right) = O\left(k\beta^2_{max}\sqrt{\log(k)/n_c}\right)\,.$$

Next, by Proposition 3.10, with probability at least $1 - 2k^{-1}$ we have $\max_{j\in 1:k}|\mathbf{X}_{v,j}^T\epsilon_v n_v^{-1}| = O(\sigma\sqrt{\log(k)/n_v})$. Therefore, again with probability at least $1 - 4k^{-1}$,

$$\frac{\epsilon_v^T\mathbf{X}_v}{n_v}(\beta - \hat{\beta}_{J_h}) = O\left(\sigma|\beta_{max}|k\sqrt{\log(k)/n_v}\right)\,.$$

Finally, with probability at least $1 - 2(2/e)^k$, we have $\max_h \|(\mathbf{X}_{c,J_h}^T \mathbf{X}_{c,J_h})^{-1}\| = O(n_c^{-1})$, so with this probability we also have

$$\begin{aligned}\frac{\epsilon_c^T P_h \epsilon_c}{n_c} &\leq \|n_c^{-1}\epsilon_c^T \mathbf{X}_{c,J_h}\| \cdot \|n_c(\mathbf{X}_{c,J_h}^T \mathbf{X}_{c,J_h})^{-1}\| \cdot \|n_c^{-1}\mathbf{X}_{c,J_h}^T \epsilon_c\| \\ &= O\left(\sigma\sqrt{k\log(k)/n_c}\right) \cdot O(1) \cdot O\left(\sigma\sqrt{k\log(k)/n_c}\right) \\ &= O\left(\sigma^2 k\log(k)/n_c\right) \ .\end{aligned}$$

Putting it all together, let $\gamma_2(k) \equiv 8k^{-1} + 2(2/e)^k$. Then with probability at least $1 - \gamma_2(k)$, we have that

$$\max_h |\nu_{J_h}| = O\left(k\beta_{max}^2\sqrt{\log(k)/n_c}\right) + O\left(\sigma|\beta_{max}|k\sqrt{\log(k)/n_v}\right) + O\left(\sigma^2 k\log(k)/n_c\right) \ .$$

Since $\beta_{min}^2/\sigma^2 \geq c \cdot k\log(k)/n_c$, our probability of underfit goes to zero if $\exists\ c' > 0$ s.t.

$$\frac{\beta_{min}^2}{\beta_{max}^2} \geq c' \cdot \max\left\{k\sqrt{\frac{\log(k)}{n_c}}, \frac{k^2\log(k)/n_v}{\beta_{min}^2/\sigma^2}\right\} \ .$$

8.1.7 Proof of Proposition 4.6

In this proof we work only with the training data. The subscript c is omitted for brevity.

Recall that X_h contains just the single column for spurious predictor h, not all columns in the spurious model $J_h = J_* \cup h$.

We make a training mistake if, using the observed construction dataset, a fitted spurious model would do better in expectation on validation data than the fitted true model, i.e. if $B_h \equiv \mathbb{E}_v\left(\widehat{MSE}(J_h) - \widehat{MSE}(J_*)\right) < 0$ for some h. Note that

$$\begin{aligned}B_h &= (\hat\beta_{J_h} - \beta)^T\Sigma(\hat\beta_{J_h} - \beta) - (\hat\beta_{J_*} - \beta)^T\Sigma(\hat\beta_{J_*} - \beta) \\ &= \left((\hat\beta_{J_*} - \hat\beta_{J_h}) - 2(\hat\beta_{J_*} - \beta)\right)^T \Sigma(\hat\beta_{J_*} - \hat\beta_{J_h})\,.\end{aligned}$$

We can write

$$\begin{aligned}\hat\beta_{J_h} = (\mathbf{X}_{J_h}^T\mathbf{X}_{J_h})^{-1}\mathbf{X}_{J_h}^T Y &= \begin{bmatrix} \mathbf{X}_*^T\mathbf{X}_* & \mathbf{X}_*^T X_h \\ X_h^T\mathbf{X}_* & X_h^T X_h \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_*^T Y \\ X_h^T Y \end{bmatrix} \\ &= \begin{bmatrix} \hat\beta_{J_*} \\ 0 \end{bmatrix} - \begin{bmatrix} (\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_*^T X_h \\ -1 \end{bmatrix} \cdot \frac{X_h^T P_*^\perp Y}{X_h^T P_*^\perp X_h}\end{aligned}$$

where the last equality is by blockwise matrix inversion, and where $P_*^\perp = I - \mathbf{X}_*(\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_*^T$.

Note that $P_*^\perp Y = P_*^\perp(\mathbf{X}_*\beta + \epsilon) = P_*^\perp \epsilon$, since $P_*^\perp \mathbf{X}_* = 0$. So let us denote the scalar fraction above as $\tilde{\beta}_{J_h} = \frac{X_h^T P_*^\perp \epsilon}{X_h^T P_*^\perp X_h}$.

Then

$$\begin{aligned}
B_h =& \left(\tilde{\beta}_{J_h}^2 \cdot \begin{bmatrix} (\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T X_h \\ -1 \end{bmatrix} - 2\tilde{\beta}_{J_h} \cdot \begin{bmatrix} (\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T \epsilon \\ 0 \end{bmatrix} \right)^T \Sigma \begin{bmatrix} (\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T X_h \\ -1 \end{bmatrix} \\
=& \left(\tilde{\beta}_{J_h}^2 \cdot \begin{bmatrix} (\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T X_h \\ -1 \end{bmatrix} - 2\tilde{\beta}_{J_h} \cdot \begin{bmatrix} (\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T \epsilon \\ 0 \end{bmatrix} \right)^T \begin{bmatrix} \Sigma_*(\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T X_h - \Sigma_{*,h} \\ -\Sigma_{*,h}^T(\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T X_h + 1 \end{bmatrix} \\
=& \tilde{\beta}_{J_h}^2 \cdot \left(X_h^T \mathbf{X}_*(\mathbf{X}_*^T \mathbf{X}_*)^{-1}\Sigma_*(\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T X_h - 2X_h^T \mathbf{X}_*(\mathbf{X}_*^T \mathbf{X}_*)^{-1}\Sigma_{*,h} + 1 \right) \\
& - 2\tilde{\beta}_{J_h} \cdot \left(\epsilon^T \mathbf{X}_*(\mathbf{X}_*^T \mathbf{X}_*)^{-1}\Sigma_*(\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T X_h - \epsilon^T \mathbf{X}_*(\mathbf{X}_*^T \mathbf{X}_*)^{-1}\Sigma_{*,h} \right) .
\end{aligned}$$

Now let $\hat{\alpha}_{X_h} = (\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T X_h$ and $\hat{\alpha}_\epsilon = (\mathbf{X}_*^T \mathbf{X}_*)^{-1}\mathbf{X}_*^T \epsilon$. Let the population versions be $\alpha_{X_h} = \Sigma_*^{-1}\Sigma_{*,h}$ and $\alpha_\epsilon = 0$ (since we assume the noise is uncorrelated with all predictors).

Then we simplify B_h above:

$$\begin{aligned}
B_h &= \tilde{\beta}_{J_h}^2 \cdot (\hat{\alpha}_{X_h}^T \Sigma_* \hat{\alpha}_{X_h} - 2\hat{\alpha}_{X_h}^T \Sigma_{*,h} + 1) - 2\tilde{\beta}_{J_h} \cdot \left(\hat{\alpha}_\epsilon^T \Sigma_* \hat{\alpha}_{X_h} - \hat{\alpha}_\epsilon^T \Sigma_{*,h} \right) \\
&= \tilde{\beta}_{J_h}^2 \cdot \left((1 - \hat{\alpha}_{X_h}^T \Sigma_{*,h}) + (\hat{\alpha}_{X_h}^T(\Sigma_* \hat{\alpha}_{X_h} - \Sigma_{*,h})) \right) - 2\tilde{\beta}_{J_h} \cdot \hat{\alpha}_\epsilon^T (\Sigma_* \hat{\alpha}_{X_h} - \Sigma_{*,h}) .
\end{aligned}$$

Note that $1 - \hat{\alpha}_{X_h}^T \Sigma_{*,h} = 1 - \alpha_{X_h}^T \Sigma_{*,h} - (\hat{\alpha}_{X_h} - \alpha_{X_h})^T \Sigma_{*,h} = 1 - \Sigma_{h,*}\Sigma_*^{-1}\Sigma_{*,h} - (\hat{\alpha}_{X_h} - \alpha_{X_h})^T \Sigma_{*,h}$, where $\gamma_{J_h} \equiv 1 - \Sigma_{h,*}\Sigma_*^{-1}\Sigma_{*,h}$ takes on a value between 0 and 1 (by properties of Schur complement).

Also, $\Sigma_* \hat{\alpha}_{X_h} - \Sigma_{*,h} = \Sigma_*(\alpha_{X_h} + \hat{\alpha}_{X_h} - \alpha_{X_h}) - \Sigma_{*,h} = \Sigma_*(\hat{\alpha}_{X_h} - \alpha_{X_h})$. Therefore,

$$\begin{aligned}
B_h &= \tilde{\beta}_{J_h}^2 \cdot \left(\gamma_{J_h} - (\hat{\alpha}_{X_h} - \alpha_{X_h})^T \Sigma_{*,h} + \hat{\alpha}_{X_h}^T \Sigma_*(\hat{\alpha}_{X_h} - \alpha_{X_h}) \right) - 2\tilde{\beta}_{J_h} \cdot \hat{\alpha}_\epsilon^T \Sigma_*(\hat{\alpha}_{X_h} - \alpha_{X_h}) \\
&= \tilde{\beta}_{J_h}^2 \cdot \left(\gamma_{J_h} + (\hat{\alpha}_{X_h} - \alpha_{X_h})^T \Sigma_*(\hat{\alpha}_{X_h} - \alpha_{X_h}) \right) - 2\tilde{\beta}_{J_h} \cdot \hat{\alpha}_\epsilon^T \Sigma_*(\hat{\alpha}_{X_h} - \alpha_{X_h})
\end{aligned}$$

for some $\gamma_{J_h} \in (0,1)$. Also, $n_c \tilde{\beta}_{J_h}^2 \approx \chi_1^2$ if $\mathbf{X}$ and ϵ are Gaussian (or we apply Lemma 8.6 if they are sub-Gaussian). And let W be the event that the following conditions hold, which happens with probability at least $1 - cp^{-1}$:

- $\max_h (\hat{\alpha}_{X_h} - \alpha_{X_h})^T \Sigma_*(\hat{\alpha}_{X_h} - \alpha_{X_h}) = O(k\log(p)/n_c)$, and
- $\max_h \hat{\alpha}_\epsilon^T \Sigma_*(\hat{\alpha}_{X_h} - \alpha_{X_h}) = O(\sigma k\sqrt{\log(p)}/n_c)$.

Then for n_c large enough,

$$\begin{aligned}
\mathbb{P}(\text{mistake on any } h) &\le \mathbb{P}(\neg W) + \mathbb{P}(\min_h B_h < 0 | W) \\
&\le cp^{-1} + p \cdot \mathbb{P}\left(\left|\tilde{\beta}_{J_h}\right| \cdot (c + O(k \log(p)/n_c)) < 2 \cdot \text{sign}\left(\tilde{\beta}_{J_h}\right) \cdot O(\sigma k \sqrt{\log(p)}/n_c) \right) \\
&\le cp^{-1} + p \cdot \mathbb{P}\left(\sqrt{n_c} \left|\tilde{\beta}_{J_h}\right| / \sigma < O(k \sqrt{\log(p)/n_c}) \right) \\
&\le cp^{-1} + c' kp \sqrt{\frac{\log(p)}{n_c}} + c'' \frac{p}{\sqrt{n_c}}
\end{aligned}$$

where the last line is by the anti-concentration result in Lemma 8.6.

(Note that if we did not cancel $\tilde{\beta}_{J_h}$ in the second line above, then the third and fourth lines would be on the order of $p \cdot \mathbb{P}\left(\chi_1^2 < O\left(n_c^{-1/2}\right)\right) \approx c' p n_c^{-1/4}$. This would require a worse rate of $p^4/n_c \to 0$ to achieve consistency, instead of only $p^2/n_c \to 0$.)

Therefore, $\mathbb{P}(\text{any mistake across } h) \to 0$ as long as $\frac{k^2 p^2 \log(p)}{n_c} \to 0$. This sufficient condition is not too far from the necessary condition that $p^2/n_c \to 0$ from Proposition 4.8.

8.1.8 Proof of Proposition 4.7

We continue on from the proof of Proposition 4.6, but now we will also account for finite testing data. We will use the c and v subscripts for training and testing sets respectively.

We cannot make a mistake unless at least one spurious model J_h gives

$$\begin{aligned}
0 &> \widehat{MSE}(J_h) - \widehat{MSE}(J_*) \\
&= 2\frac{\epsilon_v^T \mathbf{X}_{v,J_h}}{n_v}\left(\hat{\beta}_{J_*} - \hat{\beta}_{J_h}\right) + B_h \\
&\quad + \left[\left(\hat{\beta}_{J_h} - \beta\right)^T \left(\frac{\mathbf{X}_{v,J_h}^T \mathbf{X}_{v,J_h}}{n_v} - \Sigma_{J_h} \right) \left(\hat{\beta}_{J_h} - \beta\right) - \left(\hat{\beta}_{J_*} - \beta\right)^T \left(\frac{\mathbf{X}_{v,J_h}^T \mathbf{X}_{v,J_h}}{n_v} - \Sigma_{J_h} \right) \left(\hat{\beta}_{J_*} - \beta\right) \right] \\
&\in 2\frac{\epsilon_v^T \mathbf{X}_{v,J_h}}{n_v}\left(\hat{\beta}_{J_*} - \hat{\beta}_{J_h}\right) + B_h \left(1 \pm \left\| \frac{\mathbf{X}_{v,J_h}^T \mathbf{X}_{v,J_h}}{n_v} - \Sigma_{J_h} \right\| / \|\Sigma_{J_h}\| \right).
\end{aligned}$$

Recall that

$$\hat{\beta}_{J_*} - \hat{\beta}_{J_h} = \tilde{\beta}_{J_h} \cdot \begin{bmatrix} (\mathbf{X}_{c,*}^T \mathbf{X}_{c,*})^{-1} \mathbf{X}_{c,*}^T X_{c,h} \\ -1 \end{bmatrix}.$$

Then

$$2\frac{\epsilon^T \mathbf{X}_{v,J_h}}{n_v}\left(\hat{\beta}_{J_*} - \hat{\beta}_{J_h}\right) = 2\tilde{\beta}_{J_h} \cdot \left[\frac{\epsilon_v^T \mathbf{X}_{v,*}}{n_v} (\mathbf{X}_{c,*}^T \mathbf{X}_{c,*})^{-1} \mathbf{X}_{c,*}^T X_{c,h} - \frac{\epsilon_v^T X_{v,h}}{n_v} \right].$$

Let W' be the event that the following conditions hold, as well as the conditions of event W from the proof of Proposition 4.6; all this happens with probability at least $1 - cp^{-1}$:

- $\max_h \left\| \frac{\epsilon_v^T \mathbf{X}_{v,*}}{n_v} \right\| = O(\sigma\sqrt{k\log(k)/n_v})$,
- $\max_h \left\|(\mathbf{X}_{c,*}^T\mathbf{X}_{c,*})^{-1}\mathbf{X}_{c,*}^T X_{c,h}\right\| \leq \max_h \left\|(\mathbf{X}_{c,*}^T\mathbf{X}_{c,*})^{-1}\mathbf{X}_{c,*}^T X_{c,h} - \Sigma_*^{-1}\Sigma_{*,h}\right\| + \|\Sigma_*^{-1}\Sigma_{*,h}\| = O\left(\sqrt{\frac{k\log(p)}{n_c}} + \frac{1}{\sqrt{k}}\right)$,
- $\max_h \left\| \frac{\epsilon_v^T X_{v,h}}{n_v} \right\| = O(\sigma\sqrt{k\log(p)/n_v})$, and
- $\max_h \|n_v^{-1}\mathbf{X}_{v,J_h}^T\mathbf{X}_{v,J_h} - \Sigma_{J_h}\| = O\left(\|\Sigma_{J_h}\|\sqrt{\log(p)/n_v}\right)$.

Then, across all h, for n_c large enough,

$$\begin{aligned}
\mathbb{P}(\text{mistake}) &\leq \mathbb{P}(\neg W') + \mathbb{P}\left(\min_h \left(\widehat{MSE}(J_h) - \widehat{MSE}(J_*)\right) < 0 \,\middle|\, W'\right) \\
&\leq cp^{-1} + p \cdot \mathbb{P}\left[B_h \cdot \left(1 - O\left(\sqrt{\frac{\log(p)}{n_v}}\right)\right) < \right. \\
&\qquad\qquad \left. 2\sigma\tilde{\beta}_{J_h} \cdot O\left(\frac{k\log(p)}{\sqrt{n_c n_v}} + \sqrt{\frac{\log(k)}{n_v}} + \sqrt{\frac{k\log(p)}{n_v}}\right)\right] \\
&\leq cp^{-1} + p \cdot \mathbb{P}\left[\frac{\left|\tilde{\beta}_{J_h}\right|}{\sigma} < O\left(\frac{k\sqrt{\log(p)}}{n_c}\right) + O\left(\frac{k\log(p)}{\sqrt{n_c n_v}} + \sqrt{\frac{k\log(p)}{n_v}}\right)\right] \\
&\leq cp^{-1} + p \cdot \mathbb{P}\left[\frac{\sqrt{n_c}\left|\tilde{\beta}_{J_h}\right|}{\sigma} < O\left(\frac{k\sqrt{\log(p)}}{\sqrt{n_c}} + \frac{k\log(p)}{\sqrt{n_v}} + \sqrt{\frac{n_c k\log(p)}{n_v}}\right)\right] \\
&\leq cp^{-1} + c'\left(\frac{kp\sqrt{\log(p)}}{\sqrt{n_c}} + \frac{kp\log(p)}{\sqrt{n_v}} + \sqrt{\frac{n_c kp^2\log(p)}{n_v}}\right) + c''\frac{p}{\sqrt{n_c}}
\end{aligned}$$

where the last line is by the anti-concentration result in Lemma 8.6.

Therefore, $\mathbb{P}(\text{any mistake across } h) \to 0$ as long as:

- $\frac{k^2p^2\log(p)}{n_c} \to 0$, same as in Proposition 4.6;
- $\frac{n_c kp^2\log(p)}{n_v} \to 0$, which is stronger than Shao or Zhang's $n_c/n_v \to 0$ because now that ratio must go to 0 faster than p^2 grows, which is the price we pay for using the union bound across a growing model set; and
- $\frac{k^2p^2(\log(p))^2}{n_v} \to 0$, which is implied by the previous two conditions.

Again, these sufficient conditions are not too far from the necessary conditions that $p^2/n_c \to 0$ from Theorem 4.4 and that $n_c/n_v \to 0$ from Shao or Zhang's fixed-path, fixed-p setting. Compared to Shao and Zhang, we pay a price for choosing from among p models: analogously to Theorem 4.4, this price is essentially that $\frac{p^2}{n_v/n_c} \to 0$ as $n_v/n_c \to \infty$.

8.1.9 Proof of Proposition 4.8

Use vector notation. Let $\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$ and $\|x\|^2 = \langle x, x\rangle$.

Under Assumptions 7 and 8, we train the true intercept-only model $J_* = \emptyset$ and a given spurious univariate model $J_h = \{h\}$. At a new test-data observation $X_{i,h}$, the prediction from the estimated true model is $\overline{Y}_c = \mu + \overline{\epsilon}_c$, and the prediction from the estimated spurious model is $\hat{\beta}_{0,h} + \hat{\beta}_{1,h} X_{i,h} = \overline{Y}_c + \hat{\beta}_{1,h}(X_{i,h} - \overline{X}_{c,h})$, where $\overline{X}_{c,h}$, $\overline{\epsilon}_c$, and $\hat{\beta}_{1,h} = \frac{\overline{X\epsilon}_{c,h} - \overline{X}_{c,h}\overline{\epsilon}_c}{\overline{X^2}_{c,h} - \overline{X}^2_{c,h}}$ are all estimates from the training data for predictor h. In this simple case, $\overline{\epsilon}$, $\overline{X}$, and $\hat{\beta}_1$ are all mutually independent $N(0, n_c^{-1})$.

Then the true model's risk is $\mathbb{E}_v(\overline{Y}_c - \mu)^2 = \overline{\epsilon}_c^2$, and the wrong model's risk is

$$\begin{aligned}
\mathbb{E}_v \left(\overline{Y}_c - \mu + \hat{\beta}_{1,h}(X_{i,h} - \overline{X}_{c,h})\right)^2 &= \overline{\epsilon}_c^2 + \hat{\beta}_{1,h}^2 \left(\mathbb{E}_v(X_{i,h}^2) + \overline{X}_{c,h}^2 - 2\overline{X}_{c,h}\mathbb{E}_v(X_{i,h})\right) \\
&\quad + 2\overline{\epsilon}_c\hat{\beta}_{1,h}\left(\mathbb{E}_v(X_{i,h}) - \overline{X}_{c,h}\right) \\
&= \overline{\epsilon}_c^2 + \hat{\beta}_{1,h}^2\left(1 + \overline{X}_{c,h}^2\right) - 2\hat{\beta}_{1,h}\overline{X}_{c,h}\overline{\epsilon}_c
\end{aligned}$$

where $\mathbb{E}_v$ is the expectation taken over validation datasets.

We can define the difference in risks

$$B_h \equiv \hat{\beta}_{1,h}^2(1 + \overline{X}_{c,h}^2) - 2\hat{\beta}_{1,h}\overline{X}_{c,h}\overline{\epsilon}_c$$

and we will say we make a "training mistake" if $B_h < 0$ for at least one h.

(Since B_h depends only on the training data, the rest of this proof omits subscripts h and c for succinct notation, except on B_h as needed. n below actually refers to n_c, the number of training records.)

Using conditional independence of B_h given ϵ, the probability of no training mistake is

$$\begin{aligned}
\mathbb{P}\left(\min_h B_h \geq 0\right) &= \mathbb{E}\left[\mathbb{P}\left(B_h \geq 0 \ \forall h \mid \epsilon\right)\right] \\
&= \mathbb{E}\left\{\left[\mathbb{P}\left(B \geq 0 \mid \epsilon\right)\right]^p\right\} .
\end{aligned}$$

Now we consider $\mathbb{P}(B > 0 \mid \epsilon)$. For notational simplicity we drop the conditioning on ϵ and h. That is, in the following math display we consider fixed ϵ and h.

Let $Z = \langle X, \epsilon - \overline{\epsilon}\rangle/\sqrt{n}$ and $S = \sqrt{n}\overline{X}\overline{\epsilon}$. Then conditional on ϵ, Z and S are independent with distributions

$$Z \sim N\left(0, \|\epsilon - \overline{\epsilon}\|^2/n\right), \quad S \sim N(0, \overline{\epsilon}^2) .$$

Let $\frac{1+\overline{X}^2}{\|X - \overline{X}\|^2/n} = 1 + R$, where R is a function of X satisfying $P(R \geq c\sqrt{\log n/n}) \leq n^{-1}$ for some absolute constant $c > 0$. Also assume that ϵ satisfies $\frac{1}{\sqrt{n}} \leq \frac{\sqrt{n}|\overline{\epsilon}|}{\|\epsilon - \overline{\epsilon}\|} \leq 1/2$.

By cancelling a $\hat{\beta}_1$ out of B, we see that

$$\begin{aligned}
&\mathbb{P}\left(B \geq 0\right) \\
=&\mathbb{P}\left(\frac{|\langle X, \epsilon - \overline{\epsilon}\rangle|}{\|X - \overline{X}\|^2}(1 + \overline{X}^2) - 2\ \mathrm{sign}(\langle X, \epsilon - \overline{\epsilon}\rangle)\overline{X}\overline{\epsilon} \geq 0\right) \\
=&\mathbb{P}\left(|Z|(1 + R) - 2\ \mathrm{sign}(Z)S \geq 0\right) \\
=&\frac{1}{2}\mathbb{P}\left(Z(1 + R) - 2S \geq 0 \mid Z \geq 0\right) + \frac{1}{2}\mathbb{P}\left(-Z(1 + R) + 2S \geq 0 \mid Z < 0\right) \\
=&\mathbb{P}\left(Z(1 + R) - 2S \geq 0 \mid Z \geq 0\right) \quad \textit{(the two probabilities are the same by considering } X \leftarrow -X\textit{)} \\
\leq&\mathbb{P}\left(R > c\sqrt{\log n/n} \mid Z \geq 0\right) + \mathbb{P}\left(Z(1 + R) - 2S \geq 0,\ R \leq c\sqrt{\log n/n} \mid Z \geq 0\right) \\
\leq&\mathbb{P}\left(R > c\sqrt{\log n/n}\right) + \mathbb{P}\left(Z(1 + c\sqrt{\log n/n}) - 2S \geq 0 \mid Z \geq 0\right) \quad \textit{(Z > 0 and R are independent)} \\
\leq&n^{-1} + \frac{1}{2} + \frac{1}{2}\mathbb{P}\left(Z \geq \frac{2}{1 + c\sqrt{\log n/n}}S \;\middle|\; S > 0, Z \geq 0\right) \\
=&n^{-1} + \frac{1}{2} + \frac{1}{2}\mathbb{P}\left(\frac{Z/(\|\epsilon - \overline{\epsilon}\|/\sqrt{n})}{S/|\overline{\epsilon}|} \geq \frac{2}{1 + c\sqrt{\log n/n}}\frac{\sqrt{n}|\overline{\epsilon}|}{\|\epsilon - \overline{\epsilon}\|} \;\middle|\; S > 0, Z \geq 0\right) \\
=&n^{-1} + \frac{1}{2} + \frac{1}{2} - \frac{1}{\pi}\arctan\left(\frac{2}{1 + c\sqrt{\log n/n}}\frac{\sqrt{n}|\overline{\epsilon}|}{\|\epsilon - \overline{\epsilon}\|}\right) \quad \textit{(Cauchy distribution)} \\
\leq&1 + n^{-1} - \frac{1}{\pi}\arctan\left(\frac{2}{(1 + c\sqrt{\log n/n})\sqrt{n}}\right) \\
\leq&1 + n^{-1} - \frac{1}{\pi}\frac{\pi}{2(1 + c\sqrt{\log n/n})\sqrt{n}}\,.
\end{aligned}$$

Recalling Assumption 9, we have $\liminf p^2/n = \Gamma$ for some $\Gamma > 0$. Thus for such ϵ we have, for n large enough,

$$\limsup_{n\to\infty}[\mathbb{P}(B \geq 0)]^p \leq \limsup_{n\to\infty}\left[1 - \frac{\sqrt{\Gamma}/2}{p}\right]^p \leq e^{-\sqrt{\Gamma}/2}\,.$$

Let

$$A = \left\{\epsilon : 1/\sqrt{n} \leq \frac{\sqrt{n}|\overline{\epsilon}|}{\|\epsilon - \overline{\epsilon}\|} \leq 1/2\right\}\,.$$

Then by independence between $\overline{\epsilon}$ and $\epsilon - \overline{\epsilon}$, let T_{n-1} be a random variable with student t_{n-1}-distribution.

$$\mathbb{P}(\epsilon \in A) = \mathbb{P}\left(|T_{n-1}| \in [1, \sqrt{n}/2]\right) \to 2(1 - \Phi(1))\,.$$

Then

$$
\begin{aligned}
&\mathbb{P}(\min_h B_h \geq 0) \\
=&\mathbb{E}\left\{[\mathbb{P}(B \geq 0 \mid \epsilon)]^p\right\} \\
=&\mathbb{E}\left\{\mathbf{1}_A(\epsilon)\,[\mathbb{P}(B \geq 0 \mid \epsilon)]^p\right\} + \mathbb{E}\left\{\mathbf{1}_{A^c}(\epsilon)\,[\mathbb{P}(B \geq 0 \mid \epsilon)]^p\right\} \\
\leq&\sup_{\epsilon \in A}[\mathbb{P}(B \geq 0 \mid \epsilon)]^p\,\mathbb{P}(\epsilon \in A) + \mathbb{P}(\epsilon \in A^c) \\
=&1 - \mathbb{P}(\epsilon \in A)\left\{1 - \sup_{\epsilon \in A}[\mathbb{P}(B \geq 0 \mid \epsilon)]^p\right\}.
\end{aligned}
$$

Taking lim sup we have

$$
\limsup_{n\to\infty} \mathbb{P}(\min_h B_h \geq 0) \leq 1 - 2(1-\Phi(1))(1 - e^{-\sqrt{\Gamma}/2}) \leq 1 - 0.32(1 - e^{-\sqrt{\Gamma}/2}).
$$

For instance, if $p^2 \equiv n$ so that $\Gamma = 1$, then

$$
\limsup_{n\to\infty} \mathbb{P}(\min_h B_h \geq 0) \leq 1 - 0.32(1 - e^{-1/2}) \leq 1 - 0.12.
$$

The probability of a training mistake cannot vanish unless $\Gamma = 0$.

8.1.10 Proof of Theorem 4.4

We continue on from the proof of Proposition 4.8, still omitting subscript h except as needed. $\overline{X}_c$, $\bar{\epsilon}_c$, etc. are still computed on the training data, while the individual cases X_i and ϵ_i will refer to test-data records.

Here n still refers to the training sample size. The argument is uniform over all testing sample sizes n_v.

Consider $p = \sqrt{n}$. The argument can be easily extended to $p = c\sqrt{n}$ for constants $0 < c < 1$. For larger values of p, just consider the first $\sqrt{n}$ columns of $\mathbf{X}$. This way we do not need to worry about the difference between $\sqrt{\log p}$ and $\sqrt{\log n}$.

Recall that

$$
\hat{\beta} = \begin{pmatrix} \mu \\ 0 \end{pmatrix} + \begin{pmatrix} \bar{\epsilon}_c \\ 0 \end{pmatrix} + \hat{\beta}_1 \begin{pmatrix} -\overline{X}_c \\ 1 \end{pmatrix}
$$

where

$$
\hat{\beta}_1 = \frac{\overline{X\epsilon}_c - \overline{X}_c\bar{\epsilon}_c}{\overline{X^2}_c - \overline{X}_c^2}
$$

is estimated from the training data.

Then for each test observation $i \in 1, \ldots, n_v$, the difference in squared errors between the correct model and incorrect model J_h (subscript omitted) is

$$\begin{aligned}
&(\mu + \epsilon_i - \overline{Y}_c)^2 - (\mu + \epsilon_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\
= {}& \epsilon_i^2 + 2\epsilon_i(\mu - \overline{Y}_c) + (\mu - \overline{Y}_c)^2 - \epsilon_i^2 - \hat{\beta}_1^2 X_i^2 - (\mu - \hat{\beta}_0)^2 - 2\epsilon_i(\mu - \hat{\beta}_0) + 2\hat{\beta}_1\epsilon_i X_i + 2(\mu - \hat{\beta}_0)\hat{\beta}_1 X_i \\
= {}& (\mu - \overline{Y}_c)^2 - \left[(\mu - \hat{\beta}_0)^2 + \hat{\beta}_1^2\right] + 2\hat{\beta}_1\epsilon_i X_i + 2(\mu - \hat{\beta}_0)\hat{\beta}_1 X_i - \hat{\beta}_1^2(X_i^2 - 1) + 2\epsilon_i(\hat{\beta}_0 - \overline{Y}_c) \\
= {}& -B_h + 2\hat{\beta}_1\epsilon_i X_i + 2(\mu - \hat{\beta}_0)\hat{\beta}_1 X_i - \hat{\beta}_1^2(X_i^2 - 1) - 2\epsilon_i\hat{\beta}_1\overline{X}_c \\
= {}& -B_h + 2\hat{\beta}_1\left[\epsilon_i X_i + (\mu - \hat{\beta}_0)X_i - \hat{\beta}_1(X_i^2 - 1)/2 - \overline{X}_c\epsilon_i\right] \qquad (8.3)
\end{aligned}$$

where we make a model-selection mistake if the sum of these differences is positive over the test dataset, so the true model appears to have higher test MSE than the spurious model. Recall B_h from the proof of Proposition 4.8:

$$B_h \equiv \hat{\beta}_1^2(1 + \overline{X}_c^2) - 2\hat{\beta}_1\overline{X}_c\bar{\epsilon}_c = -\left((\mu - \overline{Y}_c)^2 - \left[(\mu - \hat{\beta}_0)^2 + \hat{\beta}_1^2\right]\right).$$

We want to show there is a nonvanishing probability that the $\epsilon_i X_i$ term dominates the other terms in the square brackets in Equation 8.3 while $\hat{\beta}_1\epsilon_i X_i > 0$ and $B_h \leq 0$, which leads to a model-selection mistake.

Let $W_i = (X_i^2 - 1)/2$. Define $\widetilde{X\epsilon} = \frac{1}{\sqrt{n_v}}\sum_{i=1}^{n_v} X_i\epsilon_i$, and $\widetilde{X}$, $\widetilde{\epsilon}$, $\widetilde{W}$ correspondingly.

Let event Q be such that

$$\begin{aligned}
&\sup_h \max\left\{|\mu - \hat{\beta}_0|, |\hat{\beta}_1|, |\overline{X}_c|\right\} \leq c\sqrt{\log n/n} \\
&|\bar{\epsilon}_c| \leq c\sqrt{\log n/n} \\
&\sup_h \max\left\{|\widetilde{X}|, |\widetilde{W}|\right\} \leq c\sqrt{\log n} \\
&|\widetilde{\epsilon}| \leq c\sqrt{\log n}.
\end{aligned}$$

For some absolute constant c, we have $\mathbb{P}(Q) \geq 1 - n^{-1}$.

Let event L_h be such that (note that L_h depends on h)

$$B_h \leq 0, \quad \left|\widetilde{\epsilon X}\right| \geq \frac{3c^2\log n}{\sqrt{n}}, \quad \mathrm{sign}(\widetilde{\epsilon \mathrm{X}}) = \mathrm{sign}(\hat{\beta}_1).$$

By independence between $\hat{\beta}_1$ and $\widetilde{\epsilon X}$, and the symmetry of $\widetilde{\epsilon X}$ we have

$$\mathbb{P}(L_h | B_h \leq 0) = \frac{1}{2}\mathbb{P}\left(|\widetilde{\epsilon X}| \geq \frac{3c^2\log n}{\sqrt{n}}\right).$$

When $n_v = 1$, then $\widetilde{\epsilon X} = \epsilon X$ and

$$\mathbb{P}(|\epsilon X| \geq t) = 1 - \mathbb{P}(|\epsilon X| < t) \geq 1 - \mathbb{P}(|\epsilon| < \sqrt{t}) - \mathbb{P}(|X| < \sqrt{t}) \geq 1 - c'\sqrt{t},$$

where c' is an absolute constant.

When $n_v \geq 2$, then we can write $\widetilde{\epsilon X}$ as $\frac{1}{2\sqrt{n_v}}(U - V)$ where U, V are independent $\chi^2_{n_v}$ random variables. In particular $U = \sum_{i=1}^{n_v}(\epsilon_i + X_i)^2/2$, $V = \sum_{i=1}^{n_v}(\epsilon_i - X_i)^2/2$.

Using Lemma 8.7, the density of $\widetilde{\epsilon X}$ is uniformly bounded for all $n_v \geq 2$, so there exists a constant c' such that for all $n_v \geq 2$ and all $t > 0$.

$$\mathbb{P}(\widetilde{\epsilon X} \geq t) = 1 - \mathbb{P}(|\widetilde{\epsilon X}| < t) \geq 1 - c't.$$

Now let

$$t = 3c^2 \log n/\sqrt{n}\,.$$

For n large enough, we have $t \leq 1$, and hence for some c' and uniformly over n_v

$$\mathbb{P}(|\widetilde{\epsilon X}| \geq t) \geq 1 - c'\sqrt{t}\,.$$

Now let $L = \bigcup_h L_h$, and $H = \{h : B_h \leq 0\}$. Then

$$\mathbb{P}(L) = \mathbb{P}(L,\ H \neq \emptyset) = \mathbb{P}(L|H \neq \emptyset)\mathbb{P}(H \neq \emptyset) \geq \frac{1}{2}(1 - c'\sqrt{t})\mathbb{P}(H \neq \emptyset)\,.$$

Then

$$\begin{aligned}\mathbb{P}(\text{mistake}) \geq& \mathbb{P}(L \cap Q) \geq \mathbb{P}(L) - \mathbb{P}(\neg Q) \\ \geq& \frac{1}{2}(1 - c'\sqrt{t})\mathbb{P}(H \neq \emptyset) - n^{-1}\,.\end{aligned}$$

So

$$\liminf_{n\to\infty} \mathbb{P}(\text{mistake}) \geq \frac{1}{2} \liminf_{n\to\infty} \mathbb{P}(H \neq \emptyset) \geq 0.16(1 - e^{-\sqrt{\Gamma}/2})\,.$$

For instance, if $p^2 \equiv n$ so that $\Gamma = 1$, then

$$\liminf_{n\to\infty} \mathbb{P}(\text{mistake}) \geq 0.06\,.$$

The probability of an overall model-selection mistake cannot vanish unless $\Gamma = 0$.

8.1.11 Derivation of Corollary 5.1

Under Assumptions A through D of Zhang (1993), for MCV and RLT, Theorems 1 and 4 of Zhang show that for a correct or overfitting model $J_h \supseteq J_*$, with size $h \geq k$,

$$\widehat{MSE}(J_h) = n^{-1}\epsilon^T P_{J_h}^{\perp}\epsilon + \left(1 + \frac{n}{n_c}\right) \cdot \frac{h\sigma^2}{n} + o_p(n^{-1})$$

while for underfitting $J_h \subset J_*$, with size $h < k$,

$$\widehat{MSE}(J_h) = n^{-1}\epsilon^T\epsilon + b_{J_h} + o_p(1)$$

where $P_{J_h} = \mathbf{X}_{J_h}(\mathbf{X}_{J_h}^T\mathbf{X}_{J_h})^{-1}\mathbf{X}_{J_h}^T$ and $b_{J_h} = \liminf_{n\to\infty} n^{-1}(\mathbf{X}\beta)^T P_{J_h}^{\perp}\mathbf{X}\beta$. Note that $b_1 \geq \ldots \geq b_{k-1} \geq b_k = 0$ for a path of nested submodels of J_*. For $h < k$, $\frac{b_{J_h}}{\sigma^2/n}$ is a kind of signal-to-noise ratio.

These results still hold if we modify parts of Zhang's Assumption's A and C, replacing $n^{-1}(\mathbf{X}\beta)^T P_{J_h}\mathbf{X}\beta \to 0$ with the following pair of conditions:
$\lambda \to 1$, and $\limsup_{n\to\infty} n^{-1}(\mathbf{X}\beta)^T P_{J_h}\mathbf{X}\beta = c_{J_h} < c$ for some $c < \infty$.

By an intermediate step in Zhang's own proof of Theorem 1, for $h < k$ we have

$$\begin{aligned}\widehat{MSE}(J_h) &= n^{-1}\epsilon^T P_{J_h}^{\perp}\epsilon + n^{-1}(\mathbf{X}\beta)^T P_{J_h}^{\perp}\mathbf{X}\beta + 2n^{-1}\epsilon^T P_{J_h}^{\perp}\mathbf{X}\beta + O\left(\left[n_v\binom{n}{n_v}\right]^{-1} o_p(1)\right)\\ &= n^{-1}\epsilon^T P_{J_h}^{\perp}\epsilon + b_{J_h} + 2n^{-1}\epsilon^T P_{J_h}^{\perp}\mathbf{X}\beta + o_p(n^{-1})\,.\end{aligned}$$

Since ϵ is Gaussian, the last lines's 1st term is a scaled Chi-square and the 3rd term is a scaled Gaussian plus another $o_p(n^{-1})$ term.

Now, the difference between the true model and a too-small model of size $h < k$ is

$$\frac{n}{\sigma^2} \cdot \left(\widehat{MSE}(J_h) - \widehat{MSE}(J_*)\right) = \sigma^{-2}\epsilon^T(P_* - P_{J_h})\epsilon + \frac{b_{J_h}}{\sigma^2/n} + 2\sigma^{-2}\epsilon^T P_{J_h}^{\perp}\mathbf{X}\beta - k\left(1 + \frac{n}{n_c}\right) + o_p(1)\,.$$

Note that $\sigma^{-2}\epsilon^T(P_* - P_{J_h})\epsilon \sim \chi^2_{k-h}$; and $-\sigma^{-2}\epsilon^T P_{J_h}^{\perp}\mathbf{X}\beta \sim N(0, \frac{b_{J_h}}{\sigma^2/n}) + o_p(1)$.

Then we have

$$\mathbb{P}(\text{correctly choose } J_* \text{ over } J_h) = \mathbb{P}\left(A_1 > 2A_2 + k\left(1 + \frac{n}{n_c}\right) - \frac{b_{J_h}}{\sigma^2/n} + o_p(1)\right)$$

where $A_1 \sim \chi^2_{k-h}$ and $A_2 \sim N(0, \frac{b_{J_h}}{\sigma^2/n})$ are not independent.

If we additionally assume that the $o_p(1)$ term is small enough to ignore (perhaps for sample sizes larger than some sufficiently large N), this probability becomes approximately

$$\mathbb{P}\left(\chi^2_{k-h} > 2\sqrt{\frac{b_{J_h}}{\sigma^2/n}}N(0,1) + k\left(1+\frac{n}{n_c}\right) - \frac{b_{J_h}}{\sigma^2/n}\right).$$

Let $\chi^2_{(h),\alpha}$ be the lower α quantile of χ^2_h, and let $Z_{1-\alpha}$ be the upper α quantile of $N(0,1)$. Also let $r = n_c/n$.

If we can tolerate a probability of α of making a mistake on this comparison, we can control the chi-square and Normal terms jointly at level α with a Bonferroni correction by using their $\alpha/2$ quantiles. We need to satisfy

$$\chi^2_{(k-h),\alpha/2} \geq 2\sqrt{\frac{b_{J_h}}{\sigma^2/n}}Z_{1-\frac{\alpha}{2}} + k(1+r^{-1}) - \frac{b_{J_h}}{\sigma^2/n}$$

which is a quadratic in $\sqrt{n}$:

$$n \cdot \frac{b_{J_h}}{\sigma^2} - \sqrt{n}\cdot 2\frac{\sqrt{b_{J_h}}}{\sigma}Z_{1-\frac{\alpha}{2}} + \chi^2_{(k-h),\alpha/2} - k(1+r^{-1}) \geq 0\,.$$

Using the quadratic formula, the smallest n that achieves this must satisfy

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

(The operation before the radical was ±, but we chose + instead of − to get a positive $\sqrt{n}$, because $k(1+r^{-1}) > \chi^2_{(k-h),\alpha/2}$ for any reasonably small α and thus the radical term is greater than $Z_{1-\frac{\alpha}{2}}$.)

For any n too small to satisfy this inequality at the largest possible $r \approx 1$, LOO is the best we can do. But if n is large enough to satisfy this inequality for $r = 1$, then we can start to make r smaller while retaining the same $1-\alpha$ probability of avoiding underfit. We can choose any

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

8.2 Lemmas

Lemma 8.1. *Let C be a coherence matrix: $C = \mathbf{x}^T\mathbf{x}$ for some $n \times p$ matrix $\mathbf{x}$ whose columns have unit norm, so C is symmetric with diagonal entries of 1 and off-diagonal entries' absolute values ≤ 1. Let A^TA be the Cholesky decomposition of C. Let $\gamma, \mu > 0$.*

Case 1: The greatest absolute off-diagonal entry is μ. *Then the off-diagonal entries of A are upper-bounded by $\frac{\mu}{1-(p-1)\mu}$, and the bottom-right entry is lower-bounded by $\sqrt{\frac{1-(p-1)\mu}{1-(p-2)\mu}}$.*

If we also assume that each row of C is s-sparse off of the diagonals (has s nonzero off-diagonal entries) with $1 \leq s < p$, then each off-diagonal row sum is upper-bounded as $\operatorname{rowsum}_j(|A - I|) \leq \frac{s\mu}{1-s\mu}$ *for* $j \in 1, \ldots, p$.

Case 2: μ is the greatest absolute off-diagonal entry except in the last column and row, where γ is the greatest absolute off-diagonal entry. *Then the off-diagonal entries in the last column and row of A are upper-bounded by* $\frac{\gamma}{1-(p-2)\mu-(p-1)\gamma^2}$.

We can also assume that each row of $C_{1:(p-1),1:(p-1)}$ *is s-sparse off of the diagonals with* $1 \leq s < (p-1)$. *That is, all but the last row and column of $C - I$ are s-sparse. If so, then the off-diagonal entries in the last column and row of A are upper-bounded by* $\frac{\gamma}{1-s\mu-(p-1)\gamma^2}$.

Proof. Let $E = C - I$, which has zero diagonal and bounded off-diagonal entries. Apply Theorem 2.1 of Sun (1992), which tells us that $|A - I|$ is entrywise upper-bounded by $(I - |E|)^{-1}|E|$, where $|E|$ is taking absolute values entrywise. Then

$$(I - |E|)^{-1}|E| = (I - |E|)^{-1}(I - (I - |E|)) = (I - |E|)^{-1} - I = \sum_{i=1}^{\infty} |E|^i$$

where the last equality comes from the geometric series for matrices: $(I - B)^{-1} = \sum_{i=0}^{\infty} B^i$ as long as $\|B\|_{op} < 1$.

Case 1: For $i = 1$, $|E|$ is entrywise bounded by μ. For $i = 2$, entries of $|E|^2$ are at most $\langle (0, \mu, \ldots, \mu), (0, \mu, \ldots, \mu) \rangle = (p-1)\mu^2$. For $i = 3$, entries of $|E|^3 = |E|\,|E|^2$ are at most $\langle (0, \mu, \ldots, \mu), (0, (p-1)\mu^2, \ldots, (p-1)\mu^2) \rangle = (p-1)^2\mu^3$, and so on.

By induction, $|E|^i$ is entrywise upper-bounded by $(p-1)^{i-1}\mu^i$, so $\sum_{i=1}^{\infty} |E|^i$ is entrywise upper-bounded by $\frac{\mu}{1-(p-1)\mu}$. Therefore this is an entrywise upper-bound on $|A - I|$. Off-diagonal entries of A are upper-bounded by $\frac{\mu}{1-(p-1)\mu}$, and diagonal entries of A are upper-bounded by $1 + \frac{\mu}{1-(p-1)\mu}$.

For the bottom-right entry we can also give a lower bound. Note that also

$$C^{-1} = (I - (I - C))^{-1} - I = \sum_{i=1}^{\infty} (I - C)^i \leq \sum_{i=1}^{\infty} |E|^i$$

where the last inequality is entrywise, so that $1 + \frac{\mu}{1-(p-1)\mu}$ upper-bounds the diagonal entries of C^{-1} too. Now note that the Schur complement on the bottom-right entry of C is $a_{p,p}^2 = x_p^T x_p - x_p^T \mathbf{x}_{1:p-1}(\mathbf{x}_{1:p-1}^T \mathbf{x}_{1:p-1})^{-1}\mathbf{x}_{1:p-1}^T x_p$. So $a_{p,p}^{-2}$ equals the correponding entry of C^{-1}, whose diagonals we have just bounded:

$$a_{p,p} \geq \sqrt{\frac{1}{1 + \frac{\mu}{1-(p-1)\mu}}} = \sqrt{\frac{1 - (p-1)\mu}{1 - (p-2)\mu}}.$$

Finally, now assume $C - I$ is s-sparse. Let $(e_{j,1}^{(i)}, \ldots, e_{j,p}^{(i)})$ be the jth row of $|E|^i$. Each element in this row is the inner product of the jth row of $|E|^{i-1}$ with a column of $|E| = |C - I|$. The rows of $|E|$ are all s-sparse, so every $e_{j,k}^{(i-1)}$ has a nonzero coefficient at most s times when we form the $e_{j,k}^{(i)}$. This means

$$\text{rowsum}_j \left(|E|^i\right) \le \sum_{k=1}^{p} e_{j,k}^{(i-1)} \cdot \mu \cdot s = \mu s \cdot \text{rowsum}_j(|E|^{i-1}) \, .$$

So we can write

$$\text{rowsum}_j \left(|A - I|\right) \le \text{rowsum}_j \left(\sum_{i=1}^{\infty} |E|^i\right) = \sum_{i=1}^{\infty} \text{rowsum}_j \left(|E|^i\right) = \sum_{i=1}^{\infty} (s\mu)^i = \frac{s\mu}{1 - s\mu}$$

assuming $s \le p - 1$.

Case 2: For $i = 1$, the last column of $|E|$ is entrywise bounded by γ.

For $i = 2$, the last column of $|E|^2$ is at most $\langle (0, \mu, \ldots, \mu, \gamma), (\gamma, \ldots, \gamma, 0) \rangle = (p-2)\mu\gamma$, except in the diagonal position which is at most $\langle (\gamma, \ldots, \gamma, 0), (\gamma, \ldots, \gamma, 0) \rangle = (p-1)\gamma^2$.

For $i = 3$ with $|E|^3 = |E| \; |E|^2$, and so on, we see that the new off-diagonal is at most $(p-2)\mu$ times the previous off-diagonal plus γ times the previous diagonal; and the new diagonal is at most $(p-1)\gamma$ times the previous off-diagonal.

Write it as a recurrence relation. For the final column of $|E|^j$, let the maximal off-diagonal entry be F_{j-1} and the maximal diagonal entry be G_{j-1}. We have $F_0 = \gamma$ and $G_0 = 0$, as well as $F_1 = (p-2)\mu\gamma$. Then

$$\begin{aligned} F_{j+1} &= (p-2)\mu F_j + \gamma G_j \\ G_{j+1} &= (p-1)\gamma F_j \end{aligned}$$

so we can eliminate G_j and just work with

$$F_{j+2} = (p-2)\mu F_{j+1} + (p-1)\gamma^2 F_j$$

for $j \ge 0$. We don't need a closed-form solution for F_j, just its infinite sum:

$$\sum_{j=0}^{\infty} F_{j+2} = (p-2)\mu \sum_{j=0}^{\infty} F_{j+1} + (p-1)\gamma^2 \sum_{j=0}^{\infty} F_j$$

$$\left(-\gamma - (p-2)\mu\gamma + \sum_{j=0}^{\infty} F_j\right) = (p-2)\mu \left(-\gamma + \sum_{j=0}^{\infty} F_j\right) + (p-1)\gamma^2 \sum_{j=0}^{\infty} F_j$$

$$\sum_{j=0}^{\infty} F_j = \frac{\gamma}{1 - (p-2)\mu - (p-1)\gamma^2} \, .$$

This upper-bounds the off-diagonal entries in the last column of $|A - I|$, so it upper-bounds the absolute off-diagonal entries in the last column of A.

Finally, if we also assume the sparsity condition, the desired result follows from the same proof as above after redefining $F_1 = s\mu\gamma$ in the recurrence relation. □

Lemma 8.2. $\sqrt{\frac{1-t\mu}{1-(t-1)\mu}} - \frac{(2k-2t-1)\mu}{1-(t+1)\mu}$ *is increasing in* t *for* $0 < \mu < (2k-1)^{-1}$, $t \geq 2$, $k \geq 3$.

Proof. We claim that the derivative of this quantity with respect to t is positive:

$$\frac{1}{2} \cdot \sqrt{\frac{1-(t-1)\mu}{1-t\mu}} \cdot \frac{-\mu^2}{(1-(t-1)\mu)^2} - \frac{-2\mu + (2k+1)\mu^2}{(1-(t+1)\mu)^2} \overset{?}{>} 0$$

$$\frac{1}{2} \cdot \sqrt{\frac{1-(t-1)\mu}{1-t\mu}} \cdot \frac{\mu}{(1-(t-1)\mu)^2} \overset{?}{<} \frac{2-(2k+1)\mu}{(1-(t+1)\mu)^2} \,.$$

Since $(1-(t-1)\mu)^{-2} < (1-(t+1)\mu)^{-2}$, and $\sqrt{a} < a$, the following is sufficient for the above to hold:

$$\frac{1}{2} \cdot \frac{1-(t-1)\mu}{1-t\mu} \cdot \mu \overset{?}{<} 2 - (2k+1)\mu \,.$$

Since $\mu < (2k-1)^{-1}$, we have $2 - (2k+1)\mu > 2 - (2k+1)/(2k-1) = (2k-3)/(2k-1)$.

$$\frac{\mu}{2} \cdot \frac{1-(t-1)\mu}{1-t\mu} \overset{?}{<} \frac{2k-3}{2k-1}$$

$$\mu(2k-1) \cdot \frac{1-(t-1)\mu}{1-t\mu} < \frac{1-(t-1)\mu}{1-t\mu} \overset{?}{<} 4k-6 \,.$$

Since $k \geq 3$, we have $4k - 6 \geq 12 > 2$, so

$$\frac{1-(t-1)\mu}{1-t\mu} = 1 + \frac{\mu}{1-t\mu} \overset{?}{<} 2 \iff \mu \overset{?}{<} 1 - t\mu \iff \mu \overset{?}{<} \frac{1}{1+t} \,.$$

This indeed holds because $\mu < 1/(2k-1) < 1/(t+1)$. Since we found no contradictions, the original derivative was positive. □

Lemma 8.3. *Finite linear combinations of sub-Gaussian RVs are also sub-Gaussian.*

Also, let Z *be a sub-Gaussian random variable. Then there exist constants* $c_1, c_2 > 0$ *such that* $\forall\ \delta > 0$, *we have* $\mathbb{P}(|Z| > \delta) < c_1 e^{-c_2\delta^2}$.

Proof. For the first part, use the properties of norms. If $\|Z_1\|_{\psi_2} \leq c_1$ and $\|Z_2\|_{\psi_2} \leq c_2$, then $\|aZ_1 + bZ_2\|_{\psi_2} \leq ac_1 + bc_2 < \infty$.

For the second part, recall that we defined $\|Z\|_{\psi_2} = \inf\left\{C > 0 : \mathbb{E}\exp\left(|Z|^2/C^2\right) - 1 \le 1\right\}$. By this definition, $\mathbb{E}\exp(|Z|^2/\|Z\|_{\psi_2}^2) \le 2$. Therefore, by Markov's inequality, $\forall\ \delta > 0$,

$$\mathbb{P}(|Z| > \delta) = \mathbb{P}\left[\exp(|Z|^2\|Z\|_{\psi_2}^{-2}) > \exp(\delta^2\|Z\|_{\psi_2}^{-2})\right] \le \frac{\mathbb{E}\exp(|Z|^2\|Z\|_{\psi_2}^{-2})}{\exp(\delta^2\|Z\|_{\psi_2}^{-2})} \le 2\exp(-\delta^2\|Z\|_{\psi_2}^{-2})\,.$$

□

Lemma 8.4. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ be i.i.d. from a sub-Gaussian distribution with mean $\mathbb{E}\mathbf{X}_i = 0$ and covariance matrix Σ, where Σ is a correlation matrix (has 1s on the diagonal). Let $\log p \le n$.*

Let $\check{S}(t) = \frac{1}{n}\sum_{i=1}^n(\mathbf{X}_i - t)(\mathbf{X}_i - t)^T$ be the sample coherence matrix with columns centered at t. For instance, the sample covariance matrix is $S = \check{S}(\overline{\mathbf{X}}) = \frac{1}{n}\sum_{i=1}^n(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T$.

Then for any choice of $\eta > 0$, for some $c, c' > 0$ large enough and for n large enough,

$$\mathbb{P}\left(\max_{t\in T}\left\|\check{S}(t) - \Sigma\right\|_{\infty,\infty} \ge c\sqrt{\frac{\log p}{n}}\right) \le c'p^{-\eta}$$

where $T = \{\vec{0},\ \overline{\mathbf{X}},\ (\overline{\mathbf{X}}_{-p}, 0)\}$. That is, $\check{S}(t)$ may be the uncentered sample coherence; the centered sample covariance; or a hybrid in which all columns but the last are centered. (The latter will be useful in Lemma 8.5 for handling correlations between $\mathbf{X}$ and ϵ.)

Proof. By assumption, $\mathbb{E}\mathbf{X}_i\mathbf{X}_i^T = \Sigma$ and there exists a constant $\kappa > 0$ such that

$$\sup_{u\in\mathbb{S}_2^{p-1}} \|\langle\mathbf{X}_i, u\rangle\|_{\psi_2} \le \kappa\,.$$

(If Σ is not a correlation matrix, this still holds as long as its diagonals are bounded over n.)

Recall that for a random variable Z,

$$\|Z\|_{\psi_\alpha} = \left\{\inf_{c>0}\mathbb{E}e^{|Z/c|^\alpha} \le 2\right\}$$

for $\alpha \ge 1$. Directly from this definition, we see that $\|Z^2\|_{\psi_1} = \|Z\|_{\psi_2}^2$.

We follow the argument in the proof of Lemma 3.2.2 from Vu and Lei (2012):

For $a, b \in \{1, \ldots, p\}$, let $\xi_i = (\mathbf{X}_i)_a(\mathbf{X}_i)_b$, so that

$$(\check{S}(\vec{0}) - \Sigma)_{ab} = D_{ab} = \frac{1}{n}\sum_{i=1}^n(\mathbf{X}_i)_a(\mathbf{X}_i)_b - \Sigma_{ab} = \frac{1}{n}\sum_{i=1}^n(\xi_i - \mathbb{E}\xi_i)\,.$$

If we can bound $|D_{ab}|$ with high probability, then we can bound $\max_{a,b}|(\check{S}(\vec{0}) - \Sigma)_{ab}|$ with high probability by the union bound. That will also let us bound the maximum of $(S - \Sigma)_{ab} = D_{ab} - \overline{X}_a\overline{X}_b$ and of $(\check{S}((\overline{\mathbf{X}}_{-p}, 0)) -$

$\Sigma)_{ab}$. By standard arguments, $\max_{a,b}|\overline{X}_a\overline{X}_b| = O(\log(p)/n) \ll O(\sqrt{\log(p)/n})$ with probability at least $1 - cp^{\eta}$ for some $c > 0$.

Let $Y_i = \xi_i - \mathbb{E}\xi_i$, so that nD_{ab} is the sum of n independent variables Y_i. We want to apply a version of Bernstein's inequality based on van der Vaart and Wellner (1996), Lemma 2.2.11:

Sublemma 8.1. *Let $Y_1, \ldots, Y_n$ be independent, zero-mean random variables. If we have finite constants $M, v_i > 0$ and $v \geq \sum_i v_i$ that satisfy*

$$M^2\mathbb{E}\left(e^{|Y_i|/M} - 1\right) - M\mathbb{E}|Y_i| \leq v_i/2$$

then we have

$$\mathbb{P}(|Y_1 + \ldots + Y_n| > x) \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{v + Mx}\right).$$

We can simplify this a little for our purposes:

Sublemma 8.2. *Let $Y_1, \ldots, Y_n$ be as above. If we can choose constants $M \geq \max_i \|Y_i\|_{\psi_1}$ and $v = 2nM^2$ such that $M, v < \infty$, the condition of Sublemma 8.1 is satisfied.*

Furthermore, if we choose $x = Mny$ with $y = \sqrt{\frac{6\eta \log p}{n}} < 1$ and $\eta > 0$, then

$$\mathbb{P}\left(\frac{1}{n}|Y_1 + \ldots + Y_n| > M\sqrt{6\eta} \cdot \sqrt{\frac{\log p}{n}}\right) \leq 2p^{-\eta}.$$

Proof. If we choose any finite $M \geq \max_i \|Y_i\|_{\psi_1}$, then by definition of $\|Z\|_{\psi_1}$ we have that $\mathbb{E}\left(e^{|Y_i|/M} - 1\right) \leq 1$. So if we set $v_i = 2M^2$, we see that

$$M^2\mathbb{E}\left(e^{|Y_i|/M} - 1\right) - M\mathbb{E}|Y_i| \leq M^2 \cdot 1 - M\mathbb{E}|Y_i| \leq M^2 = v_i/2$$

so the condition of Sublemma 8.1 is satisfied.

Also, plug in $x = Mny$ to get

$$\mathbb{P}\left(\frac{1}{n}|Y_1 + \ldots + Y_n| > Mx\right) \leq 2e^{-\frac{1}{2}\frac{(Mny)^2}{2nM^2 + nM^2y}} = 2e^{-\frac{1}{2}\frac{ny^2}{2+y}}.$$

Finally, choose $\eta > 0$ and $0 < y = \sqrt{\frac{6\eta \log p}{n}} < 1$ to see that

$$\mathbb{P}\left(\frac{1}{n}|Y_1 + \ldots + Y_n| > My\right) \leq 2p^{-\eta\frac{3}{2+y}} \leq 2p^{-\eta}.$$

□

Since our $Y_i = \xi_i - \mathbb{E}\xi_i$ are i.i.d., we just need to upper-bound $\|\xi_i - \mathbb{E}\xi_i\|_{\psi_1}$. By van der Vaart and Wellner (1996), note that $\mathbb{E}|Z| \leq \|Z\|_{\psi_1}$ and that $\|1\|_{\psi_1} = (\log 2)^{-1}$.

By these properties, the properties of norms, and Jensen's inequality,

$$\begin{aligned}\|\xi_i - \mathbb{E}\xi_i\|_{\psi_1} &\leq \|\xi_i\|_{\psi_1} + \|\mathbb{E}\xi_i\|_{\psi_1} = \|\xi_i\|_{\psi_1} + |\mathbb{E}\xi_i| \cdot \|1\|_{\psi_1} = \|\xi_i\|_{\psi_1} + (\log 2)^{-1}|\mathbb{E}\xi_i| \\ &\leq \|\xi_i\|_{\psi_1} + (\log 2)^{-1}\mathbb{E}|\xi_i| \\ &\leq \|\xi_i\|_{\psi_1} + (\log 2)^{-1}\|\xi_i\|_{\psi_1} = \|\xi_i\|_{\psi_1}\left(1 + (\log 2)^{-1}\right) \approx 2.443\|\xi_i\|_{\psi_1} \\ &\leq 3\|\xi_i\|_{\psi_1}\,.\end{aligned}$$

Finally, we need to confirm that we can upper-bound $3\|\xi_i\|_{\psi_1}$ by a constant. Recall that $\|Z^2\|_{\psi_1} = \|Z\|_{\psi_2}^2$.

$$\begin{aligned}\|\xi_i\|_{\psi_1} &= \|(\mathbf{X}_i)_a(\mathbf{X}_i)_b\|_{\psi_1} \\ &\leq \left\|\frac{1}{2}\left((\mathbf{X}_i)_a^2 + (\mathbf{X}_i)_b^2\right)\right\|_{\psi_1} \\ &\leq \frac{1}{2}\left(\|(\mathbf{X}_i)_a^2\|_{\psi_1} + \|(\mathbf{X}_i)_b^2\|_{\psi_1}\right) = \frac{1}{2}\left(\|(\mathbf{X}_i)_a\|_{\psi_2}^2 + \|(\mathbf{X}_i)_b\|_{\psi_2}^2\right) \\ &\leq \max_{j \in 1,\ldots,p} \|\langle \mathbf{X}_i, 1_j\rangle\|_{\psi_2}^2 \\ &\leq \kappa^2\,.\end{aligned}$$

(In first inequality above: For random variables F, G with $|F(\omega)| \leq |G(\omega)|$ a.s., we have $\|F\|_{\psi_1} \leq \|G\|_{\psi_1}$. This is satisfied if $F = 2AB$ and $G = A^2 + B^2$ for random variables A, B.)

With this bound, we apply Sublemma 8.2 with $M = 3\kappa^2 \geq 3\|\xi_i\|_{\psi_1} \geq \|Y_i\|_{\psi_1}$ to say that

$$\mathbb{P}\left(|D_{ab}| > 3\kappa^2\sqrt{6\eta} \cdot \sqrt{\frac{\log p}{n}}\right) \leq 2p^{-\eta}\,.$$

Finally, using a union bound,

$$\mathbb{P}\left(\max_{a,b}|D_{ab}| > 3\kappa^2\sqrt{6(\eta+2)} \cdot \sqrt{\frac{\log p}{n}}\right) \leq p^2 \cdot 2p^{-(\eta+2)} = 2p^{-\eta}$$

and so $\max_{t\in T}\left\|\check{S}(t) - \Sigma\right\|_{\infty,\infty} = O\left(\sqrt{\frac{\log p}{n}}\right)$ with high probability. □

Lemma 8.5. *Assume the conditions of Lemma 8.4. Also assume that $\frac{\log p}{n} \to 0$. Denote the entries of the sample and population covariance matrices there (S and Σ) as s_{jk} and σ_{jk}, respectively, for $j, k \in 1, \ldots, p$. Let the sample and population correlation matrices (C and, again, Σ) have entries $r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}$ and $\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}$, respectively. (Since we assumed $\sigma_{jj} = 1$ for all j, we have $\sigma_{jk} = \rho_{jk}$ for all j, k.)*

Then we can choose $\eta > 0$ such that, for some $c, c' > 0$ large enough and for n large enough,

$$\mathbb{P}\left(\max_{j,k\in 1,\ldots,p} |r_{jk} - \rho_{jk}| \geq c\sqrt{\frac{\log p}{n}}\right) \leq c'p^{-\eta}\,.$$

Next, consider augmenting each observation $\mathbf{X}_i$ with one more variable, the noise ϵ_i. Assume the noise is uncorrelated with each predictor, so $\sigma_{j,p+1} = 0$ for all $j \in 1, \ldots, p$. Then $\hat{\gamma}\|\epsilon\|/\sqrt{n} = \max_{j\in 1,\ldots,p}\left|\frac{s_{j,p+1}}{\sqrt{s_{jj}}}\right| = O(\sigma\sqrt{\log(p)/n})$ with high probability $1 - c'p^{-\eta}$ too.

Proof. Let us assume there is a constant s.t. $0 < c < |\sigma_{jj}|$ for all $j \in 1, \ldots, p$, so $|\sigma_{jj}^{-1}| < c^{-1} < \infty$. Then with probability at least $1 - c'p^{-\eta}\ldots$

$$|s_{jj} - \sigma_{jj}| \leq c_1\sqrt{\frac{\log p}{n}} \quad\Rightarrow\quad \left|\frac{s_{jj}}{\sigma_{jj}} - 1\right| \leq c_{2j}\sqrt{\frac{\log p}{n}}$$

so that

$$\frac{s_{jj}}{\sigma_{jj}} \in \left(\max\left\{0, 1 - c_{2j}\sqrt{\frac{\log p}{n}}\right\}, 1 + c_{2j}\sqrt{\frac{\log p}{n}}\right)\,.$$

Since $\frac{\log p}{n} \to 0$, there is a large enough $N_j > 0$ such that for all $n > N_j$,

$$1 - c_{2j}\sqrt{\frac{\log p}{n}} > 1 - c_{2j}\sqrt{\frac{\log p}{N_j}} > 0$$

so for $n > N_j$,

$$\frac{s_{jj}}{\sigma_{jj}} > 1 - c_{2j}\sqrt{\frac{\log p}{n}} > 0 \quad\Rightarrow\quad \sqrt{\frac{\sigma_{jj}}{s_{jj}}} \in \left(1 \pm c_{2j}\sqrt{\frac{\log p}{n}}\right)^{-1/2} < \infty$$

and then, for $c_2 = \max\{c_{2j}, c_{2k}\}$ and $n > \max\{N_j, N_k\}$,

$$\sqrt{\frac{\sigma_{jj}\sigma_{kk}}{s_{jj}s_{kk}}} \in \left(1 \pm c_2\sqrt{\frac{\log p}{n}}\right)^{-1} \subset (0, \infty)\,.$$

Further,

$$\left|\frac{s_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}} - \rho_{jk}\right| = \left|\frac{s_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}} - \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}\right| \leq c_3\sqrt{\frac{\log p}{n}} \quad\Rightarrow\quad \frac{s_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}} \in \rho_{jk} \pm c_3\sqrt{\frac{\log p}{n}}$$

so

$$r_{jk} = \frac{s_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}} \cdot \frac{\sqrt{\sigma_{jj}\sigma_{kk}}}{\sqrt{s_{jj}s_{kk}}} \in \left(\min_{+,-}\frac{\rho_{jk} - c_3\sqrt{\frac{\log p}{n}}}{1 \pm c_2\sqrt{\frac{\log p}{n}}},\ \max_{+,-}\frac{\rho_{jk} + c_3\sqrt{\frac{\log p}{n}}}{1 \pm c_2\sqrt{\frac{\log p}{n}}}\right)$$

where the choice of $+$ or $-$ in the denominators (which are positive since $n > \max\{N_j, N_k\}$) depends on whether each numerator is positive or negative. So

$$\begin{aligned} r_{jk} - \rho_{jk} &\in \left(\min_{+,-} \frac{\rho_{jk} - c_3\sqrt{\frac{\log p}{n}}}{1 \pm c_2\sqrt{\frac{\log p}{n}}} - \rho_{jk},\ \max_{+,-} \frac{\rho_{jk} + c_3\sqrt{\frac{\log p}{n}}}{1 \pm c_2\sqrt{\frac{\log p}{n}}} - \rho_{jk} \right) \\ &\subset \left(\min_{+,-} \frac{\rho_{jk} - c_3\sqrt{\frac{\log p}{n}} - \rho_{jk}\left(1 \pm c_2\sqrt{\frac{\log p}{n}}\right)}{1 \pm c_2\sqrt{\frac{\log p}{n}}},\ \max_{+,-} \frac{\rho_{jk} + c_3\sqrt{\frac{\log p}{n}} - \rho_{jk}\left(1 \pm c_2\sqrt{\frac{\log p}{n}}\right)}{1 \pm c_2\sqrt{\frac{\log p}{n}}} \right) \\ &\subset \left(\pm \frac{(c_3 + |\rho_{jk}|c_2)\sqrt{\frac{\log p}{n}}}{1 - c_2\sqrt{\frac{\log p}{n}}} \right). \end{aligned}$$

Finally, this gives that for $n > \max\{N_j, N_k\}$

$$|r_{jk} - \rho_{jk}| \leq \frac{c_3 + |\rho_{jk}|c_2}{1 - c_2\sqrt{\frac{\log p}{\max\{N_j, N_k\}}}}\sqrt{\frac{\log p}{n}} \leq c_4\sqrt{\frac{\log p}{n}}$$

and so indeed $|r_{jk} - \rho_{jk}| = O\left(\sqrt{\frac{\log p}{n}}\right)$, with high probability. This bound holds simultaneously for each j, k with probability at least $1 - c'p^{-\eta}$, and so also for their maximum.

Additionally, we are interested in the sample coherence between standardized predictors and raw (unstandardized) noise. Let j be the index of a particular predictor, and $p+1$ be the index of the noise ϵ. We can repeat the argument above, but without dividing by $\sqrt{s_{p+1,p+1}} \equiv \sqrt{\|\epsilon\|^2/n}$. We assumed that the noise is uncorrelated with the predictors, so $\sigma_{j,p+1} = \rho_{j,p+1} = 0$ for each $j \leq p$. Omitting the $\sqrt{\frac{\sigma_{kk}}{s_{kk}}}$ factor from the derivations above, we can bound the needed entries as $\left|\frac{s_{j,p+1}}{\sqrt{s_{jj}}}\right| = O\left(\sigma\sqrt{\frac{\log p}{n}}\right)$ with the same high probability.

(To justify the extra factor of σ, note that Lemma 8.4 assumes all variances are 1 to give $|s_{j,p+1} - \sigma_{j,p+1}| = O(\sqrt{\log(p)/n})$. But the variance of each ϵ_i is σ instead of 1, which is equivalent to multiplying column $p+1$ by a constant factor of σ, which then appears inside the big-O term.) □

Lemma 8.6. *Assume 1 and 2, and let* $\log(p)/n_c \to 0$. *Define* $\tilde{\beta}_{J_h}$ *as in Section 4.3.2.*

Then, for a given h *and for* n_c *large enough,* $\sqrt{n_c}\tilde{\beta}_{J_h}/\sigma$ *exhibits anti-concentration:* $\forall\ t > 0$, *for some* $c, c', c'' > 0$,

$$\mathbb{P}\left(\sqrt{n_c}\left|\tilde{\beta}_{J_h}\right|/\sigma \leq t\right) \leq cp^{-2} + c't + c''/(\sigma^3\sqrt{n_c}).$$

Proof. We have that

$$\tilde{\beta}_{J_h} \equiv \frac{X_{c,h}^T P_*^{\perp}\epsilon_c}{X_{c,h}^T P_*^{\perp} X_{c,h}} = \frac{n_c^{-1}X_{c,h}^T\epsilon_c - n_c^{-1}X_{c,h}^T P_*\epsilon_c}{n_c^{-1}X_{c,h}^T P_*^{\perp} X_{c,h}}.$$

Note that $\forall h,\ \Sigma_{h,*}\Sigma_*^{-1}\Sigma_{*,h} \leq k \cdot \left((2k-1)^{-1}\right)^2 \cdot O(1) = O(1/k)$. Define $a_h \equiv 1 - \Sigma_{h,*}\Sigma_*^{-1}\Sigma_{*,h} = 1 + O(1/k)$, and let $A = \liminf_{n\to\infty} \min_h a_h$, where $a_h, A > 0$.

Now, $\tilde{\beta}_{J_h} \geq n_c^{-1}|X_{c,h}^T \epsilon_c| \cdot (a_h + R)$ where with probability at least $1 - c_1 p^{-2}$, $|R| \leq c_2 \cdot \sigma\sqrt{\log(p)/n_c}$ for some $c_1, c_2 > 0$, because

$$\max_h |n_c^{-1} X_{c,h}^T P_* \epsilon_c - \Sigma_{h,*}\Sigma_*^{-1}\vec{0}| = O(\sigma\sqrt{\log(p)/n_c})$$

$$\max_h |n_c^{-1} X_{c,h}^T P_*^\perp X_{c,h} - (1 - \Sigma_{h,*}\Sigma_*^{-1}\Sigma_{*,h})| = O(\sqrt{\log(p)/n_c})\,.$$

Since $X_{c,h}$ and ϵ_c are sub-Gaussian random variables, they have bounded third moments. Let us say that each is bounded by ϱ. Since each $X_{c,h}$ is independent of ϵ_c, each element of $X_{c,h}^T \epsilon_c$ also has finite third moment, at most ϱ^2. Therefore by Berry-Esseen, for all real t,

$$\left|\mathbb{P}\left(\frac{\sqrt{n_c}}{\sigma} \cdot \frac{X_{c,h}^T \epsilon_c}{n_c} \leq t\right) - \Phi(t)\right| \leq \frac{c\varrho^2}{\sigma^3\sqrt{n_c}}$$

where $\Phi(t)$ is the standard Normal CDF. Hence,

$$\mathbb{P}\left(|X_{c,h}^T \epsilon_c|/(\sigma\sqrt{n_c}) \leq t\right) \leq \Phi(t) - \Phi(-t) + 2\frac{c\varrho^2}{\sigma^3\sqrt{n_c}} \leq ct + c'/(\sigma^3\sqrt{n_c})\,.$$

For large enough n_c, we have $c_2 \cdot \sigma\sqrt{\log(p)/n_c} \leq A/2$, so that $a_h + R > A + R \geq A/2$ with high probability, and so

$$\begin{aligned}\mathbb{P}\left(\sqrt{n_c}\left|\tilde{\beta}_{J_h}\right|/\sigma \leq t\right) &\leq \mathbb{P}(|R| > A/2) + \mathbb{P}\left(|X_{c,h}^T \epsilon_c|/(\sigma\sqrt{n_c}) \leq 2t/A \mid |R| < A/2\right) \\ &\leq c_1 p^{-2} + ct + c'/(\sigma^3\sqrt{n_c})\,.\end{aligned}$$

□

Lemma 8.7. *Let U, V be independent $\chi^2_{n_v}$ random variables.*

The density of $U/\sqrt{n_v}$ is bounded uniformly for all $n_v \geq 2$. As a consequence, the density of $\frac{1}{2\sqrt{n_v}}(U - V)$ is uniformly bounded for all $n_v \geq 2$.

Proof. The first part follows directly from the density function of χ^2 distributions. The second part follows from the fact that the maximum convolution density is bounded by the individual maximum density. □

Bibliography

An, H. and Gu, L. (1985). On the selection of regression variables. *Acta Mathematicae Applicatae Sinica*, 2(1):27–36. 5

An, H., Huang, D., Yao, Q., and Zhang, C.-H. (2008). Stepwise searching for feature variables in high-dimensional linear regression. *The London School of Economics and Political Science*. 4, 5

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79. 6

Barron, A. R., Cohen, A., Dahmen, W., and DeVore, R. A. (2008). Approximation and learning by greedy algorithms. *The Annals of Statistics*, pages 64–94. 4

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, 5(Sep):1089–1105. 58

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The Million Song Dataset. In *ISMIR*, volume 2, page 10. 41, 51

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852. 4

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press. 58

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression: The X-random case. *International Statistical Review*, pages 291–319. 35

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media. 19

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

Burman, P. (1989). A comparative study of ordinary cross-validation, V-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514. 6

Burman, P. (1990). Estimation of optimal transformations using V-fold cross validation and repeated learning-testing methods. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 314–345. 6

Cai, T. T. and Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688. 2, 4, 5, 8, 19

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

Chrysostomou, K. (2009). Wrapper feature selection. In *Encyclopedia of Data Warehousing and Mining, Second Edition*, pages 2103–2108. IGI Global. 10, 28

Das, A. and Kempe, D. (2011). Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1057–1064. 11, 14, 15

Davenport, M. A. and Wakin, M. B. (2010). Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Transactions on Information Theory*, 56(9):4395–4401. 4

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7):571. 55

Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, 72(357):77–91. 3

Derksen, S. and Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282. 3

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923. 56

Donoho, D. L. and Tsaig, Y. (2008). Fast solution of norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812. 5

Draper, N. R. and Smith, H. (1966). *Applied Regression Analysis*. John Wiley & Sons, Inc, 1st edition. 2

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499. 5, 8

Efroymson, M. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers*, pages 191–203. 1

Elenberg, E. R., Khanna, R., Dimakis, A. G., and Negahban, S. (2017). Restricted strong convexity implies weak submodularity. *arXiv preprint arXiv:1612.00804v2*. 15

Fithian, W., Taylor, J., Tibshirani, R., and Tibshirani, R. (2015). Selective sequential model selection. *arXiv preprint arXiv:1512.02565*. 5

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975. 4

Genovese, C. R., Jin, J., Wasserman, L., and Yao, Z. (2012). A comparison of the lasso and marginal regression. *Journal of Machine Learning Research*, 13(Jun):2107–2143. 5

Harrell, F. (2015). *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal regression, and Survival Analysis*. Springer. 2, 55

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102. 49

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media. xiii, 49, 50

Hastie, T., Tibshirani, R., and Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*. 4

Ho, C.-H. and Lin, C.-J. (2012). Large-scale linear support vector regression. *Journal of Machine Learning Research*, 13(Nov):3323–3348. 51

Hung, K. and Fithian, W. (2016). Rank verification for exponential families. *arXiv preprint arXiv:1610.03944*. 57

Hyun, S., G'Sell, M., and Tibshirani, R. J. (2018). Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097. 9

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer. 2

Jehan, T. (2010). *Analyze Documentation*. Echo Nest Analyze API version 2.2. 51

John, G. H., Kohavi, R., Pfleger, K., et al. (1994). Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning*, pages 121–129. 6

Johnson, K. D., Stine, R. A., and Foster, D. P. (2015). Submodularity in statistics: Comparing the success of model selection methods. *arXiv preprint arXiv:1510.06301*. 11

Klein, M., Wright, T., and Wieczorek, J. (2018). A simple joint confidence region for a ranking of K populations: Application to American Community Survey's travel time to work data. *Research Report Series, U.S. Bureau of the Census*. 57

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145. 35

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324. 10

Lei, J. (2017). Cross-validation with confidence. *arXiv preprint arXiv:1703.07904*. 56

Lichman, M. (2017). UCI Machine Learning Repository. 51

Lin, D., Foster, D. P., and Ungar, L. H. (2012). VIF regression: a fast regression algorithm for large data. *Journal of the American Statistical Association*. 5

Lumley, T. based on Fortran code by Miller, A. (2017). *leaps: Regression Subset Selection*. R package version 3.0. 48

Maj-Kańska, A., Pokarowski, P., and Prochenka, A. (2015). Delete or merge regressors for linear model selection. *Electronic Journal of Statistics*, 9(2):1749–1778. 5

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462. 19

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473. 5

Nadeau, C. and Bengio, Y. (2000). Inference for the generalization error. In *Advances in Neural Information Processing Systems*, pages 307–313. 58

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming*, 14(1):265–294. 12

Pati, Y. C., Rezaiifar, R., and Krishnaprasad, P. S. (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*, pages 40–44. IEEE. 1

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 48

Revolution Analytics and Weston, S. (2015). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.10. 48

Roecker, E. B. (1991). Prediction error and its estimation for subset-selected models. *Technometrics*, 33(4):459–468. 3

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494. 6

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, pages 221–242. 5

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *The Journal of Urology*, 141(5):1076–1083. 49

Sun, J.-G. (1992). Componentwise perturbation bounds for some matrix decompositions. *BIT Numerical Mathematics*, 32(4):702–714. 17, 83

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288. 2

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620. 5

Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242. 2, 4, 8, 19

van de Geer, S. and Lederer, J. (2013). The Bernstein–Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157(1-2):225–250. 18

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer Science & Business Media. 18, 87, 88

Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*. 70

Vu, V. Q. and Lei, J. (2012). Minimax rates of estimation for sparse PCA in high dimensions. In *AISTATS*, volume 15, pages 1278–1286. 86

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524. 4

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics*, 37(5A):2178. 6

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. 48

Wiegand, R. E. (2010). Performance of using multiple stepwise algorithms for variable selection. *Statistics in Medicine*, 29(15):1647–1659. 3

Yang, Y. (2006). Comparing learning methods for classification. *Statistica Sinica*, pages 635–657. 35

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, pages 2450–2473. 6, 27

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, pages 299–313. 6, 36, 39, 81

Zhang, T. (2009). On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10(Mar):555–568. 4

Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57(7):4689–4708. xiii, 5, 49, 50

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2015). Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340. 51

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563. 19

Zheng, X. and Loh, W.-Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association*, 90(429):151–156. 5