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

Jerzy Wieczorek[1] and Jing Lei[2]

[1] *Colby College and* [2] *Carnegie Mellon University*

Supplementary material

Contents

S1 Detailed theoretical results: Path-consistency of FS, under incoherence and beta-min conditions

This appendix contains supporting results about path-consistency of FS, assuming the true model size is known. Non-trivial proofs are deferred to Appendix S4.

S1.1 Fixed design and noise

(Note: here and in the next section, 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 S1.3, we will return to random $\mathbf{X}$ with columns of unit variance instead.)

Assumption S1. 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. 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 S1. *Assume S1 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 S2. Assume S1. 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 S3. Assume S1. 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 S1, 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 S4. Assume S1. 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 S1 and Corollary S2, 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 S5. Assume S1. 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 will choose the correct model if the signal-to-noise ratio is at least

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

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 S5 is less restrictive than that of Corollary S2 only if $s \ll k$.

S1.2 Fixed design, random noise

Assumption S2. Define $\mathbf{x}$, $\{\beta_1, \ldots, \beta_k\}$, and $y = \beta_1 x_1 + \ldots + \beta_k x_k + \mathrm{E}$ as in Assumption S1, 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 S6. *Assume S2. 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 $1 - cp^{-\eta}$ for some $c > 0$).*

Corollary S7. Assume S2. 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 S6 and Corollary S2. The condition that $\sigma^2 k^2 \log(p)/n \to 0$ ensures that $\hat{\mu}$ and $\hat{\gamma}$ are both below $(2k-1)^{-1}$ for Corollary S2. □

S1.3 Random design and noise

Now use the assumptions and setup of Section 3.1, where the columns of $\mathbf{X}$ have unit variance.

Proposition S8. *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 $\hat{\mu} < (2k-1)^{-1}$.

S2 Detailed theoretical results: FS with sequential data splitting is model-selection consistent, if p grows not too fast and n_c/n_v shrinks

This appendix contains supporting results about the SeqCV-based stopping rule. Proofs are deferred to Appendix S4.

S2.1 $\mathbb{P}(underfit) \to 0$

Proposition S9. *Assume 1, 2, 4. 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*
$1 - \big(cp^{-1} + c'(k^{-1} + (2/e)^k)(1 - cp^{-1})\big) \to 1.$

S2.2 $p^2/n_c \to 0$ is sufficient for $\mathbb{P}(overfit) \to 0$

For a given training dataset and spurious covariate h, recall 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 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 will 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 S10. *Assume 1, 2, and 4. 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 S11. *Assume 1, 2, and 4. 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* $n \to \infty$*:*

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

Theorem 3.2 follows directly from the above Propositions. First, by Proposition S9, 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 S11, with probability approaching 1, model J_* will have lower $\widehat{MSE}$ than any of these $p-k$ spurious models, so FS+SeqCV must stop at model J_*.

Finally, although the Propositions are stated for the single-split version, FS+SeqCV is also model-selection consistent for any fixed number of splits of Monte Carlo CV. 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 splits. Hence, with high probability, the true model will have lowest $\widehat{MSE}$ on every split, and therefore lowest average $\widehat{MSE}$ across splits. A union bound takes care of the fact that splits on the same dataset are not independent.

S2.3 $p^2/n_c \to 0$ is necessary for $\mathbb{P}(overfit) \to 0$

Under Assumptions 5 and 6, 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

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

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

Proposition S12. *Assume 5, 6, 7.*

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.

S3 Additional numerical and simulation results

S3.1 Numerical examples of differences between OMP and FS

Let the true model be $Y = \beta_1 X_1 + \beta_2 X_2$, with no noise. Let there be three fixed predictors to choose from, with correlations $\rho_{1,2} = a$, $\rho_{1,3} = b$, and $\rho_{2,3} = c$. Assume the columns of $\mathbf{X}$ are normalized to zero mean and unit norm. Both FS and OMP will choose X_1 first if its covariance with Y is highest:

$$|Cov(X_1, Y)| = |\beta_1 + a\beta_2| > \max\{|Cov(X_2, Y)| = |a\beta_1 + \beta_2|,\ |Cov(X_3, Y)| = |b\beta_1 + c\beta_2|\}$$

Next, both algorithms will compute the residual of Y on X_1:

$$Res(Y|X_1) = Y - X_1 \cdot Cov(X_1, Y) = \beta_2(X_2 - aX_1)$$

Now, OMP will choose X_2 correctly if its covariance with this residual beats X_3's covariance with the residual.

$$|Cov(X_2, Res(Y|X_1))| = |\beta_2(1 - a^2)| > |Cov(X_3, Res(Y|X_1))| = |\beta_2(c - ab)|$$

So OMP chooses correctly if $1 - a^2 > |c - ab|$.

Meanwhile, FS must compare correlations, because the residuals $Res(X_j|X_1)$ can have different norms for $j = 2, 3$. This is equivalent to taking the same covariances used by OMP and dividing by the norm of $Res(X_j|X_1)$:

$$\|Res(X_2|X_1)\|_2 = \|Var(X_2 - aX_1)\|_2 = \sqrt{1-a^2}; \quad \|Res(X_3|X_1)\|_2 = \sqrt{1-b^2}$$

So FS chooses correctly if $\sqrt{1-a^2} > \frac{|c-ab|}{\sqrt{1-b^2}}$.

Example 1. Let the true model parameters be $\beta_1 = 2$ and $\beta_2 = 1$, with $\mathbf{X}$ correlations $a = 0.5$, $b = 0.25$, and $c = 0.9$.

Both models correctly choose X_1 first:

$$\beta_1 + a\beta_2 = 2.5 > \max\{a\beta_1 + \beta_2 \ = 2,\ b\beta_1 + c\beta_2 = 1.4\}$$

Then FS correctly chooses X_2 second:

$$\sqrt{1-a^2} \approx 0.87 > 0.80 \approx \frac{|c-ab|}{\sqrt{1-b^2}}$$

but OMP incorrectly chooses X_3 second:

$$1 - a^2 = 0.75 < 0.775 = |c - ab|$$

With $K = 2, p = 3$ it is impossible for OMP to work while FS fails. If OMP's first step is correct, so is FS's. But since FS minimizes RSS, its second and final step here must be correct too. However, at larger $p = 4$ and $K = 3$ we can find examples where OMP beats FS or vice versa. Let the true model be $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, with no noise.

Example 2. OMP selects the correct model, but FS fails on the 2nd step, if we set $\{\beta_1, \beta_2, \beta_3\} = \{5, 1.8, 1.9\}$ and $\{\rho_{1,2}, \rho_{1,3}, \rho_{2,3}, \rho_{1,4}, \rho_{2,4}, \rho_{3,4}\} = \{-0.19, 0.53, 0.16, -0.58, 0.87, 0.01\}$.

Example 3. FS selects the correct model, but OMP fails on the 2nd step, if we set $\{\beta_1, \beta_2, \beta_3\} = \{5, 1.1, 3\}$ and $\{\rho_{1,2}, \rho_{1,3}, \rho_{2,3}, \rho_{1,4}, \rho_{2,4}, \rho_{3,4}\} = \{0.76, 0.79, 0.56, -0.03, 0.5, -0.12\}$.

Using the R package `shiny` by Chang et al. (2016), we have made interactive versions of Examples 2 and 3 available online at `https://civilstat.shinyapps.io/GreedyRegrApp/`

S3.2 Tradeoffs between underfit and overfit

Section 4 discusses approximate formulas for asymptotic probabilities of underfitting or overfitting along a fixed, correct model path.

Figure S1 illustrates our approximate formulas for 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 $\frac{b_{J_{k-1}}}{\sigma^2/n}$ 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.

Next, in Figure S2 we simulate empirical estimates corresponding to the curves from Figure S1. 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.

- $\widehat{\mathbb{P}}(\text{underfit}) = \widehat{\mathbb{P}}\left(\exists\, t < k : \widehat{MSE}(J_t) < \widehat{MSE}(J_k)\right)$. The simulations indicate that our heuristic underfit curves from Equation 2 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 2.

- $\widehat{\mathbb{P}}(\text{overfit}) = \widehat{\mathbb{P}}\left(\widehat{MSE}(J_{k+1}) < \widehat{MSE}(J_k)\right)$. 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

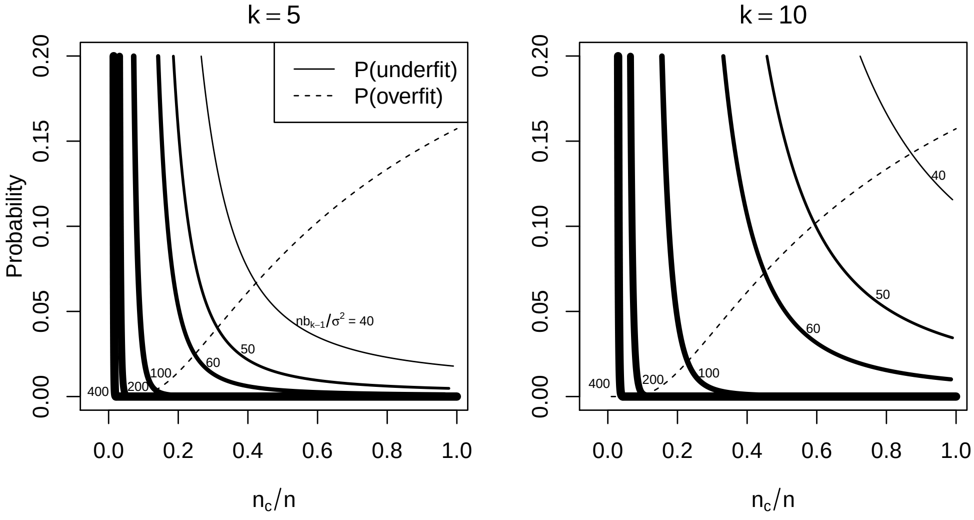

Figure S1: For $k = 5$ (left figure) and $k = 10$ (right figure), we plot the $\mathbb{P}$(overfit) (dashed line) and the $\mathbb{P}$(underfit) (solid lines at different levels of signal-to-noise ratio), vs. training ratio. Estimated using Zhang (1993) and Corollary 4.

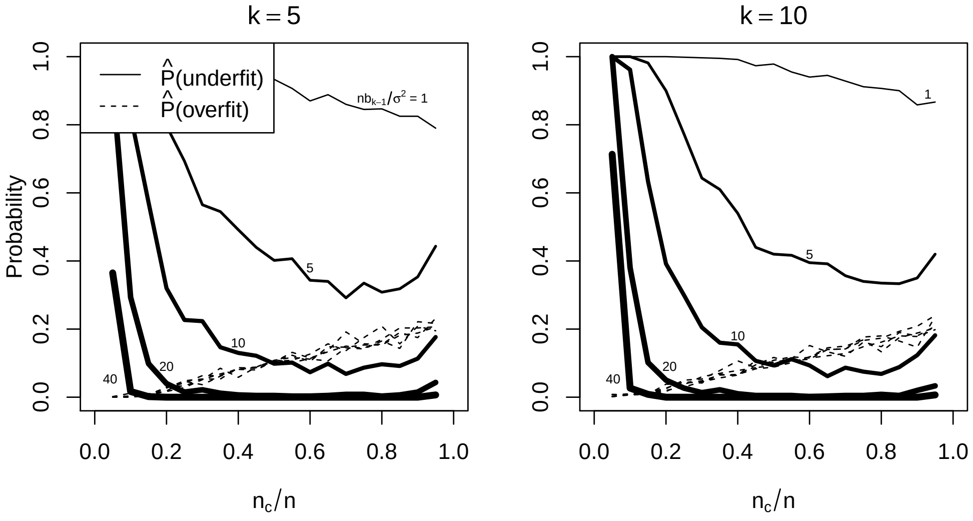

Figure S2: For $k = 5$ (left figure) and $k = 10$ (right figure), we plot the $\widehat{\mathbb{P}}$(overfit) (dashed lines) and $\widehat{\mathbb{P}}$(underfit) (solid lines), at different levels of signal-to-noise ratio, vs. training ratio. Estimated from 600 simulations at each combination of n_c/n and SNR, with orthogonal Gaussian design & noise, a fixed correct path, $n = 500$, and $p = k + 1$.

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 $\sqrt{b_{J_{k-1}}}$ in which it makes sense to tune n_c/n.

S3.3 Additional simulation results

Figures S3 and S4 use the same layout and simulation settings as Figure 2, 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 3.3, 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.

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.

Figure S4 also illustrates the detrimental effect of moderate k when $n \approx p$ as reported by Su (2018): even with a high SNR, spurious variables can be selected early when the true model is not very sparse. Consider the results shown for oracle FS for $\mu = 1/(2k)$ and $p = n = 250$. At low $k = 5$ or $k = 25$, we see less than one false positive on average, but for moderate $k = 125$ there are around 20 false positives. In other words, at extreme sparsity there is almost never a spurious variable, but at moderate sparsity the first spurious variable must be entering at least 20 steps before the correct model size is reached. On the other hand, when $p = n = 1250$, even our largest $k = 125$ is small enough to fall in the extreme sparsity regime with no early spurious variables included. These results are comparable to Su's Figure 2(a).

Finally, Figure S5 illustrates the effect of heavy-tailed noise, using a similar layout to Figure 2 but drawing super-Gaussian $\epsilon \sim t(df = 2)$ instead. (The simulations with t_2 noise use a range of smaller p and k values, which we found to adequately illustrate the difficulty of model selection at these settings.) 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 S5 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.

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 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 and datasets for reproducing our data analyses and simulations are available at `https://github.com/civilstat/FS-SeqCV-article`.

S4 Proofs

S4.1 Proof of Proposition S1

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 $p - k > 1$.

The case of $t = 0$

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 $\rho_{j,\epsilon} = \langle x_j, \text{E}/\|\text{E}\| \rangle$, 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

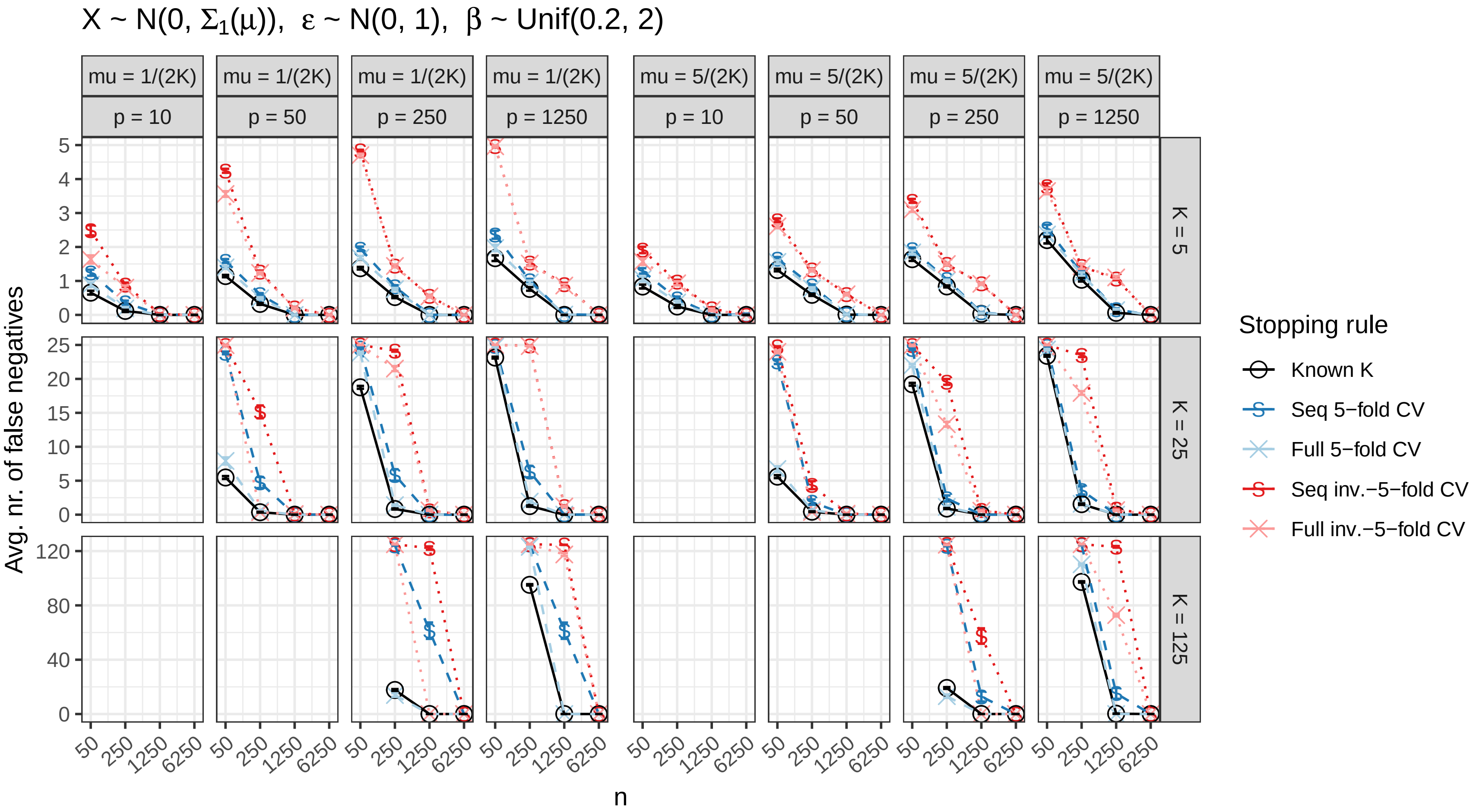

Figure S3: Average count of false negatives over 400 repetitions with constant-correlation setting $\Sigma_1(\mu)$. Incoherence condition is met in left half of figure, but not in right half. Empty subplots are impossible combinations $k > p$ or $k > n$. Error bars show $\pm 2 \cdot SE$.

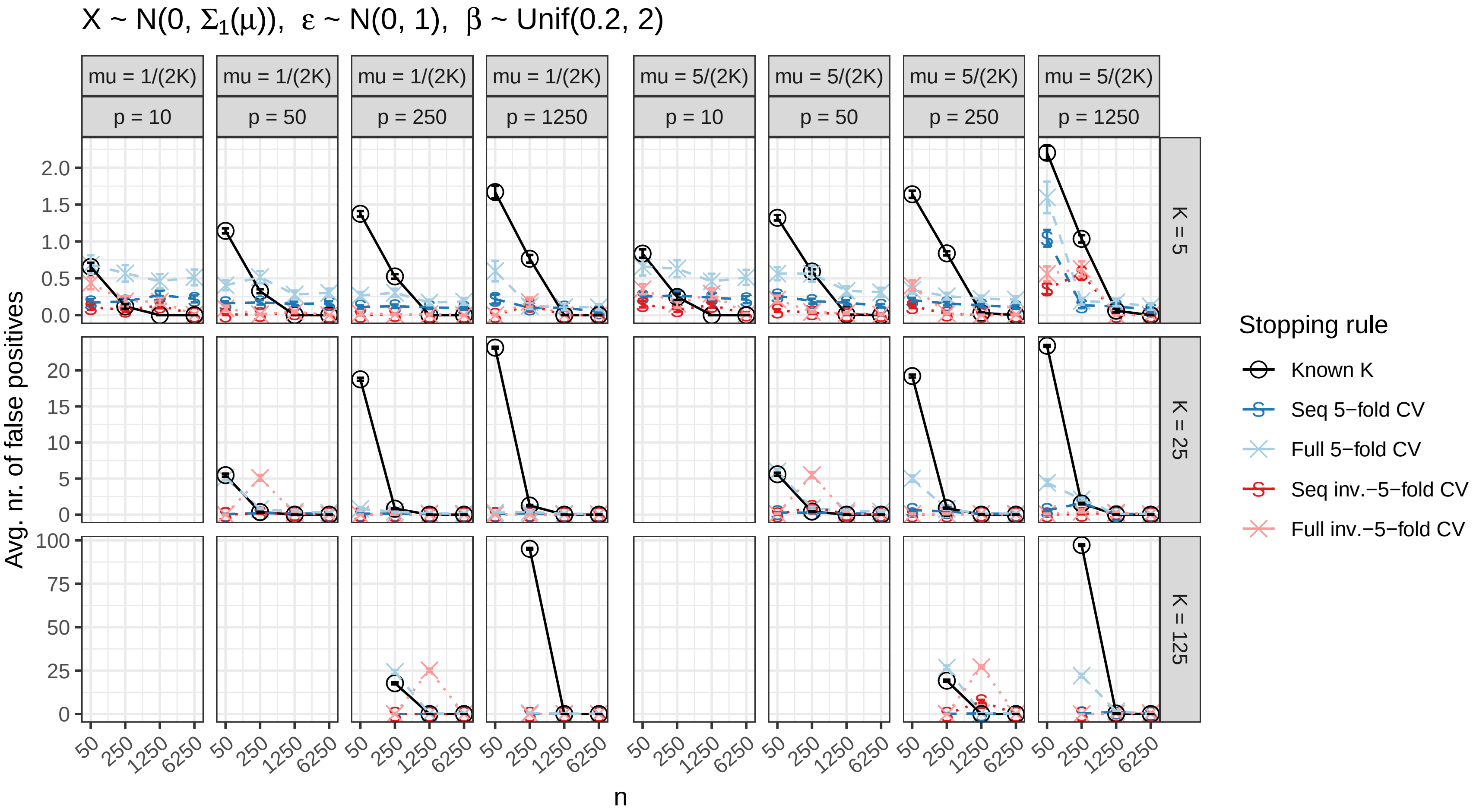

Figure S4: Average count of false positives over 400 repetitions with constant-correlation setting $\Sigma_1(\mu)$. Incoherence condition is met in left half of figure, but not in right half. Empty subplots are impossible combinations $k > p$ or $k > n$. Error bars show $\pm 2 \cdot SE$.

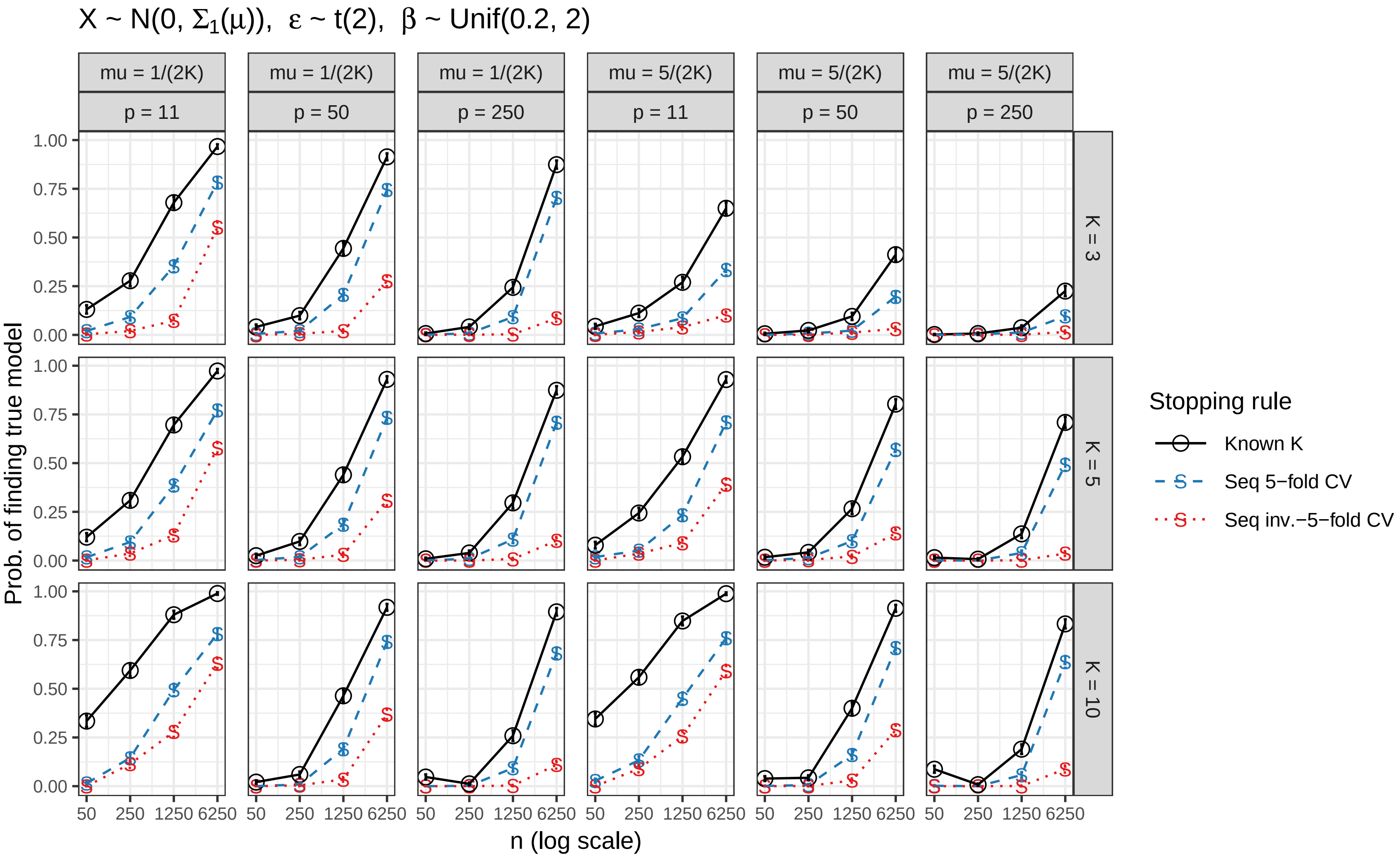

Figure S5: Proportion of correct model selections over 1000 repetitions with constant-correlation setting $\Sigma_1(\mu)$ and heavy-tailed noise $\epsilon \sim t(df = 2)$. Incoherence condition is met in left half of figure, but not in right half. Empty subplots are impossible combinations $k > p$ or $k > n$. Error bars show $\pm 2 \cdot SE$.

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

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

The case of $t = 1$

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:

$$\frac{|\beta_2|}{\|\mathrm{E}\|} > \frac{|\rho_{2,\epsilon} - \rho_{1,\epsilon}\rho_{1,2}| + |\rho_{k+1,\epsilon} - \rho_{1,\epsilon}\rho_{1,k+1}|}{1 - \mu^2 - |\rho_{k+1,2} - \rho_{1,2}\rho_{1,k+1}| - \sum_{j=3}^{k}(|\rho_{2,j} - \rho_{1,j}\rho_{1,2}| + |\rho_{k+1,j} - \rho_{1,j}\rho_{1,k+1}|)} . \quad \text{(S4.1)}$$

The RHS numerator of (S4.1) 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 (S4.1):

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

The case of $t > 1$

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 $C = A^T A$, 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 S13, 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 S13 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

$$\frac{|\beta_{t+1}|}{\|\mathrm{E}\|} > \frac{\frac{2\gamma}{1-t\mu-(t+1)\gamma^2}}{\sqrt{\frac{1-t\mu}{1-(t-1)\mu}} - \frac{(2k-2t-1)\mu}{1-(t+1)\mu}}\,. \tag{S4.2}$$

S4.2 Proof of Corollary S2

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 (S4.2) 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 S14 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

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

S4.3 Proof of Corollary S5

Repeat the proof of Proposition S1, but use the s-sparse results from Cases 1 and 2 of Lemma S13. 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 $s \to \infty$.

S4.4 Proof of Proposition S6

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 S15, 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 $1 - c_1 p^{-\eta}$.

S4.5 Proof of Proposition S8

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 S16 shows that $\|S - \Sigma\|_{\infty,\infty} = O(\sqrt{\log(p)/n})$ with high probability. Lemma S19 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 S19, 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 S19 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 $1 - (c_1 + c_2)p^{-\eta}$.

S4.6 Proof of Proposition S9

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 S1, 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 $\beta_{min}^2 > g(\beta_{max}, k, n_c, \sigma)$.

Therefore, $\beta_{min}^2 > g(\beta_{max}, k, n_c, \sigma)$ 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:

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

or

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

as $n \to \infty$. Specifically,

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

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 S1 and Lemma S13). By the QR decomposition approach in the proof of Proposition S1, 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 S13, $|a_{1,j}| \leq \frac{\mu}{1-\mu}$ for all other $j = 2, \dots, k$. 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 S13, $|a_{|J_h|+1,|J_h|+1}| \geq \sqrt{\frac{1-|J_h|\mu}{1-(|J_h|-1)\mu}}$ and $|a_{|J_h|+1,j}| \leq \frac{\mu}{1-(|J_h|+1)\mu}$ for all other $j = |J_h| + 2, \dots, k$. 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 S14, 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_{J_*} has constant off-diagonal correlation μ, we have $\Sigma_{J_*} = (1-\mu)I_k + \mu\mathbf{1}_k\mathbf{1}_k^T$, 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 $\|\Sigma_{J_*}\|, \|\Sigma_{J_*}^{-1}\| = O(1)$. These upper and lower bounds also hold for Σ_{J_h} for any sub-model $J_h \subset J_*$.

By Vershynin (2012), Theorem 5.39, let $A_{J_h} = \Sigma_{J_h}^{-1/2}\mathbf{X}_{J_h}$ which is isotropic, and then

$$\mathbb{P}\left(\|A_{J_h}\| < \sqrt{n} + c_1\sqrt{|J_h|} + t\right) \geq 1 - 2\exp(-c_2t^2)\,.$$

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:

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

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 S8, 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.

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

S4.7 Proof of Proposition S10

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

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

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:

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

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,

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

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 S20 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,

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

where the last line is by the anti-concentration result in Lemma S20.

(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 S12.

S4.8 Proof of Proposition S11

We continue on from the proof of Proposition S10, 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 S10; 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 S20.

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 S10;
- $\frac{n_c kp^2\log(p)}{n_v} \to 0$, 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 3.3 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 3.3, this price is essentially that $\frac{p^2}{n_v/n_c} \to 0$ as $n_v/n_c \to \infty$.

S4.9 Proof of Proposition S12

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 5 and 6, 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

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

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

We can define the difference in risks

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

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

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

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 $\frac{1}{\sqrt{n}} \leq \frac{\sqrt{n}|\bar{\epsilon}|}{\|\epsilon - \bar{\epsilon}\|} \leq 1/2$.

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

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

Recalling Assumption 7, 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 $\limsup$ 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 $\Gamma = 0$.

S4.10 Proof of Theorem 3.3

We continue on from the proof of Proposition S12, 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

$$\begin{aligned}
& (\mu + \epsilon_i - \overline{Y}_c)^2 - (\mu + \epsilon_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\
= {} & \epsilon_i^2 + 2\epsilon_i(\mu - \overline{Y}_c) + (\mu - \overline{Y}_c)^2 - \epsilon_i^2 - \hat{\beta}_1^2 X_i^2 - (\mu - \hat{\beta}_0)^2 - 2\epsilon_i(\mu - \hat{\beta}_0) + 2\hat{\beta}_1\epsilon_i X_i + 2(\mu - \hat{\beta}_0)\hat{\beta}_1 X_i \\
= {} & (\mu - \overline{Y}_c)^2 - \left[(\mu - \hat{\beta}_0)^2 + \hat{\beta}_1^2\right] + 2\hat{\beta}_1\epsilon_i X_i + 2(\mu - \hat{\beta}_0)\hat{\beta}_1 X_i - \hat{\beta}_1^2(X_i^2 - 1) + 2\epsilon_i(\hat{\beta}_0 - \overline{Y}_c) \\
= {} & -B_h + 2\hat{\beta}_1\epsilon_i X_i + 2(\mu - \hat{\beta}_0)\hat{\beta}_1 X_i - \hat{\beta}_1^2(X_i^2 - 1) - 2\epsilon_i\hat{\beta}_1\overline{X}_c \\
= {} & -B_h + 2\hat{\beta}_1 \left[\epsilon_i X_i + (\mu - \hat{\beta}_0)X_i - \hat{\beta}_1(X_i^2 - 1)/2 - \overline{X}_c\epsilon_i\right] \qquad \text{(S4.3)}
\end{aligned}$$

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 S12:

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

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 S4.3 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

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

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)

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

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 S21, 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

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

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

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

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

S4.11 Derivation of beta-min condition in Equation (1)

Let $Y = \mathbf{X}\beta + \epsilon$, where the columns of $\mathbf{X}$ are centered and standardized, and where $\beta_j \neq 0$ iff $j \leq k$. Let $\hat{J}_t$ be the index set of columns of $\mathbf{X}$ chosen by step t of running FS. Define the response residuals after step t as $R_t \equiv Res(Y|\mathbf{X}_{\hat{J}_t})$, and let $\tilde{R}_{t,j} \equiv Res(X_j|\mathbf{X}_{\hat{J}_t})$. Let $R_{min} \equiv \min_{j>k, J\subset\{1:k\}} Res(X_j|\mathbf{X}_J)$ and $M \equiv \max_{j>k} \|(\mathbf{X}_{1:k}^T\mathbf{X}_{1:k})^{-1}\mathbf{X}_{1:k}^T X_j\|_1$.

Whereas OMP proceeds by maximizing $|X_j^T R_t|$ over $j \notin \hat{J}_t$ at each step, FS instead maximizes the correlation between $\tilde{R}_{t,j}$ and R_t, which is equivalent to maximizing $|X_j^T R_t|/\|\tilde{R}_{t,j}\|_2$. Equation (1) arises by following the steps in Tropp (2004) and Cai and Wang (2011), who derive beta-min conditions for OMP, and modifying the approach to apply to FS instead by keeping track of terms resulting from this extra $\|\tilde{R}_{t,j}\|_2$ factor.

First, assume $\epsilon = 0$. Then by straighforwardly modifying the argument in Tropp (2004), we see that if FS has chosen correct variables up to step t, it will also choose a correct variable next if $M/R_{min} < 1$.

Next, let the noise ϵ be nonzero but fixed. If P_t is the projection matrix onto the columns of $\mathbf{X}_{\hat{J}_t}$, we can break down the residuals into signal and noise terms: $R_t = s_t + n_t$, where $s_t = (I - P_t)\mathbf{X}\beta$ and $n_t = (I - P_t)\epsilon$. Define $N_t = \max_{j\notin\hat{J}_t} \frac{|X_j^T n_t|}{\|\tilde{R}_{t,j}\|_2}$. Let $u_t = \{j : j \leq k \cap j \notin \hat{J}_t\}$ be the set of correct variables still unchosen by step t. Now, by the argument in Cai and Wang (2011), we see that if FS has a correct path so far, its next choice will also be correct if $M/R_{min} < 1$ and

$$\|\beta_{u_t}\|_2 > \frac{2\sqrt{k-t}N_t}{(1 - M/R_{min})\lambda_{min}\left(\mathbf{X}_{1:k}^T\mathbf{X}_{1:k}\right)}.$$

Finally, for sub-Gaussian noise ϵ, the arguments of Cai and Wang (2011) show that for some $c, c' > 0$, we have $N_t \leq c\sigma\sqrt{\log(p)}/R_{min}$ over all of the first k steps with probability at least

$1 - kc'/\sqrt{\log(\max\{n,p\})}$. Hence, if we have $M/R_{min} < 1$ and if for each $t < k$ we have

$$\|\beta_{u_t}\|_2 > \frac{c\sigma\sqrt{\log(p)}}{(R_{min} - M)\lambda_{min}\left(\mathbf{X}_{1:k}^T\mathbf{X}_{1:k}\right)},$$

then with high probability each of the first k steps of FS will choose a correct variable. The beta-min condition Equation (1) follows.

S4.12 Derivation of Corollary 4

Let $\epsilon \sim N(0, \sigma^2)$ and let $\mathbf{X}$ be a deterministic sequence. Assume the model path is fixed and the predictors are ordered, so that model J_h corresponds to using the first h predictors. Consider comparing true model J_* (of size k) against a particular underfitting model J_t (of size t), where $J_t \subsetneq J_*$.

Under Assumptions A through D of Zhang (1993), for MCV and MCCV, 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 $c < \infty$.

This leaves us with the following conditions:

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

B. $\sup_{n_v\to\infty} \sup_s \|n_v^{-1}\mathbf{X}_{s,J_t}^T \mathbf{X}_{s,J_t} - V_t\| = o(1)$, where $V_t, t \leq p$ 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 $t < k$, $b_{J_t} = \liminf_{n\to\infty} n^{-1}(\mathbf{X}\beta)^T P_{J_t}^{\perp} \mathbf{X}\beta > 0$ and

$c_{J_t} = \limsup_{n\to\infty} n^{-1}(\mathbf{X}\beta)^T P_{J_t} \mathbf{X}\beta < c$ for some $c < \infty$;

D. For $t \leq p$, $\max_{i\leq n} H_{ii}^{(t)} \to 0$, where $H_{ii}^{(t)}$ are the diagonal elements of P_{J_t}.

Then 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

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

Note that $\sigma^{-2}\epsilon^T(P_{J_*} - P_{J_h})\epsilon \sim \chi^2_{k-h}$; 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 $A_1 \sim \chi^2_{k-t}$; $A_2 \sim N(0, b_{J_t} n\sigma^{-2})$; and A_1 and A_2 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 $r \approx 1$, LOOCV 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

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

S5 Lemmas

Lemma S13. 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 $\text{rowsum}_j\left(|A - I|\right) \leq \frac{s\mu}{1-s\mu}$ 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

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

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:

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

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

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

So we can write

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

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

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

$$G_{j+1} = (p-1)\gamma F_j$$

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 S14. $\sqrt{\frac{1-t\mu}{1-(t-1)\mu}} - \frac{(2k-2t-1)\mu}{1-(t+1)\mu}$ 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

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

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

Lemma S15. 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 $\|Z\|_{\psi_2} = \inf\{C > 0 : \mathbb{E}\exp(|Z|^2/C^2) - 1 \leq 1\}$. By this definition, $\mathbb{E}\exp(|Z|^2/\|Z\|_{\psi_2}^2) \leq 2$. Therefore, by Markov's inequality, $\forall\ \delta > 0$,

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

$\square$

Lemma S16. 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 $\log p \leq n$.

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,

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

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 S19 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

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

(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,

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

for $\alpha \geq 1$. 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 S17. 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 S18. 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 S17 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 S17 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 S18 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 S19. Assume the conditions of Lemma S16. 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 $1 - c'p^{-\eta}$...

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

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 S16 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 S20. Assume 1 and 2, and let $\log(p)/n_c \to 0$. Define $\tilde{\beta}_{J_h}$ as in Section S2.

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 $1 - c_1p^{-2}$, $|R| \leq c_2 \cdot \sigma\sqrt{\log(p)/n_c}$ 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

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

□

Lemma S21. 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

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

Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson (2016). *shiny: Web Application Framework for R.* R package version 0.13.2.

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

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

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

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

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

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

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

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok (Eds.), *Compressed Sensing: Theory and Applications*, Chapter 5, pp. 210–268. Cambridge: Cambridge University Press.

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

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

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