Model Selection and Stopping Rules for High-Dimensional Forward Selection

Jerzy Wieczorek
Colby College

10/22/2019
Bowdoin College

This material is based on joint work with Jing Lei, supported by the NSF under Grant No. DMS-1553884.

Summary

- When can Forward Selection (FS) select the correct model?
  - When true model size is known, predictors are not too correlated or numerous, & signal is not too small
- Can Cross-Validation (CV) be used as a practical stopping rule for FS to select the correct model?
  - Yes, if the training:testing ratio decreases as sample size grows
- How should we choose the training:testing ratio?
  - We propose a rule of thumb & illustrate its use on real data

Variable selection in high-dimensional regression

Sparse linear regression: $Y = \mathbf{X}\beta + \epsilon$

- ϵ: zero-mean noise
- $\mathbf{X}$: n cases, p variables
- β: coefficient vector, with all but k 0s
- $J_* = \{j : \beta_j \neq 0\}$: "true" model, with $|J_*| = k$
- High-dim. setting: As $n \to \infty$, allow $p, k \to \infty$ and even $p \geq n$

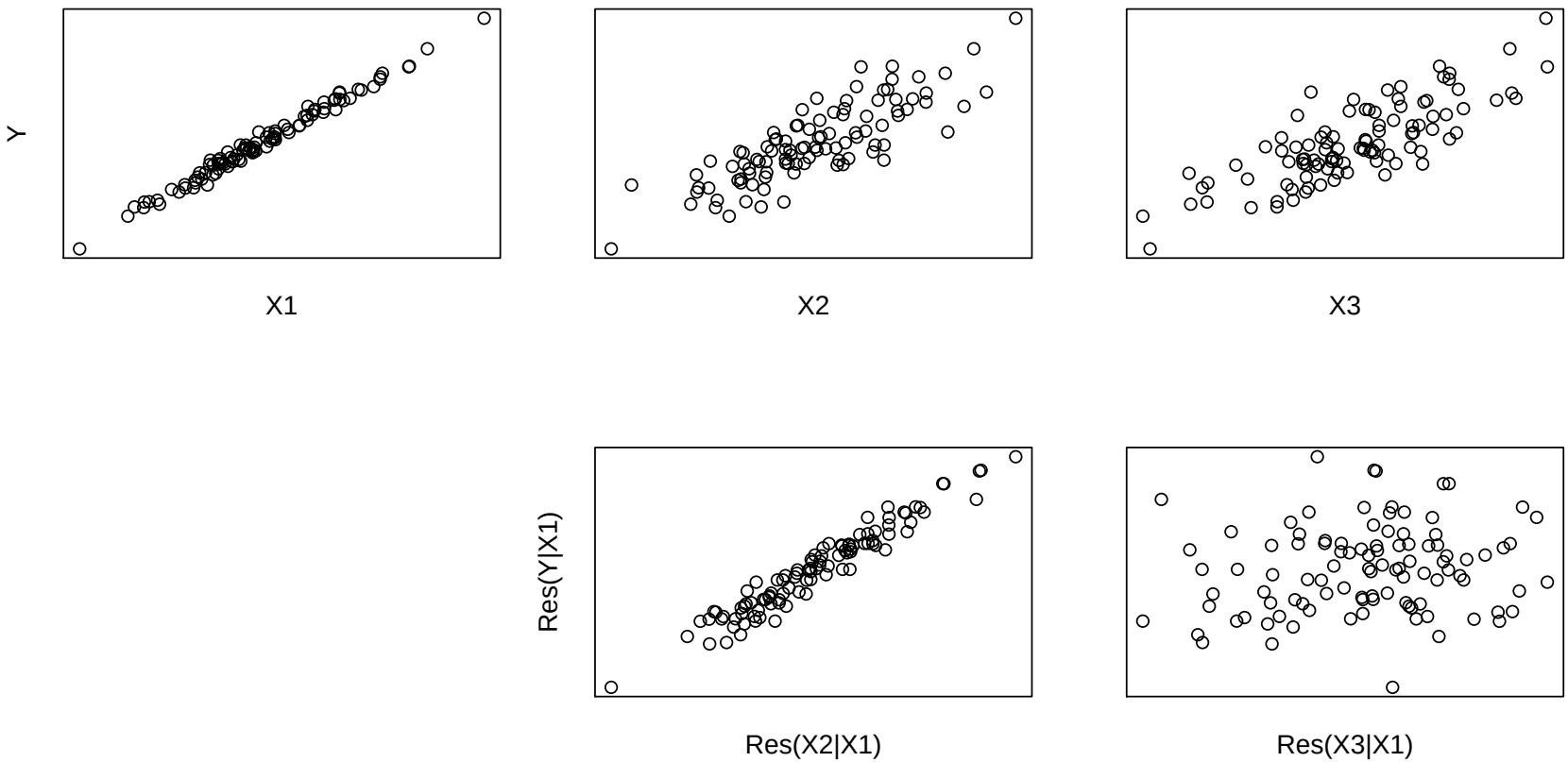
Residual Sum of Squares for some model J:

- $Res_i(Y|\mathbf{X}_J) = Y_i - \mathbf{X}_i\hat{\beta}_J$
- $RSS(Y|\mathbf{X}_J) = \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i\hat{\beta}_J\right)^2$

Forward Selection (FS)

FS algorithm (Efroymson 1960): **greedily minimize RSS**

- At step t, add "best" remaining predictor X_j:
  $\widehat{J}_{t+1} \leftarrow \widehat{J}_t \cup \arg\min_j RSS\left(Y|\mathbf{X}_{\widehat{J}_t \cup j}\right)$
- Stop at preset t or when drop in RSS is "small"; return $\widehat{J}_{FS}$



Why study FS and model selection?

- Popularity of FS
  - Simple to explain
  - Low computational cost
  - Can run even when $p > n$
- Justified criticism of FS
  - Naive inference as if final model were chosen *a priori*
  - Often too greedy under multicollinearity
  - Selected model often misinterpreted as "true"
- Value of studying FS
  - Widely taught and commonly used in practice
  - Could lead to guidance on more appropriate use
- **Model-selection consistency:**
  - $P(\widehat{J}_{FS} = J_*) \to 1$ as $n \to \infty$
  - Not the only important property, but natural to wonder about

Theorem 1: sufficient conditions for FS with k known

Assume:

- $Y = \mathbf{X}\beta + \epsilon$, where β is k-sparse with k **known** (for now)
- $Var(\epsilon_i) = \sigma^2$, and each X_j has mean 0 and variance 1
- $\mathbf{X}$ and ϵ are uncorrelated, with i.i.d. rows from some sub-Gaussian ("light-tailed") distribution

Define the "coherence" $\rho_{max} = \max_{j \neq j'} |Corr(X_j, X_{j'})|$.

Then FS is model-selection consistent if

- $\frac{\sigma^2 k^2 \log(p)}{n} \to 0$
- $|\beta_{min}| \gtrsim \sigma k \sqrt{\frac{\log(p)}{n}}$
- $\rho_{max} < \frac{1}{2k-1}$

Intuition for "coherence bound" $\rho_{max} < \frac{1}{2k-1}$

Let $Y = X_1 + \ldots + X_k$ and $p = k + 1$,
so $\epsilon \equiv 0$ and $\beta_1 = \ldots = \beta_k = 1$. Let

$$Corr(\mathbf{X}) = \begin{bmatrix} 1 & -\rho & \cdots & -\rho & \rho \\ -\rho & 1 & & -\rho & \rho \\ \vdots & & \ddots & & \vdots \\ -\rho & -\rho & & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{bmatrix}$$

If $\rho > \frac{1}{2k-1}$, then for any $i \leq k$,

$$Cov(Y, X_i) = 1 + (k-1) \cdot (-\rho) < k \cdot \rho = Cov(Y, X_{k+1})$$

so first step of FS incorrectly chooses X_{k+1}.
Thus, this bound on ρ_{max} is not improvable.

Theorem 1 proof sketch

Assume we've already correctly selected terms $J_t = \{1, \ldots, t\}$.
At step t, FS maximizes this over $j \notin J_t$:

$$f_t(j) = \left| \left\langle Res(Y|X_{J_t}), \frac{Res(X_j|X_{J_t})}{\|Res(X_j|X_{J_t})\|} \right\rangle \right|$$

- Take QR decomposition $\mathbf{X} = QR$, where Q_j are orthonormal columns, and R is upper triangular with elements $r_{i,j}$.
  Can rewrite $f_t(t+1) = \left| \sum_{j=t+1}^{k} r_{t+1,j} \beta_j \right|$.
- Also, $R^T R$ is Cholesky decomposition of $\Sigma = X^T X = R^T R$.
  Matrix perturbation approach (Sun 1992) lets us bound $|r_{t+1,j}|$ in terms of extreme off-diagonal entries of Σ, i.e. ρ_{max}.
- Find conditions on $|\beta_{min}|, \rho_{max}$ s.t. $f_t(t+1) > f_t(k+1)$.

Stopping rules when k is unknown: Cross-Validation (CV)

Popular stopping rules without strong distributional assumptions:

Sample-splitting:

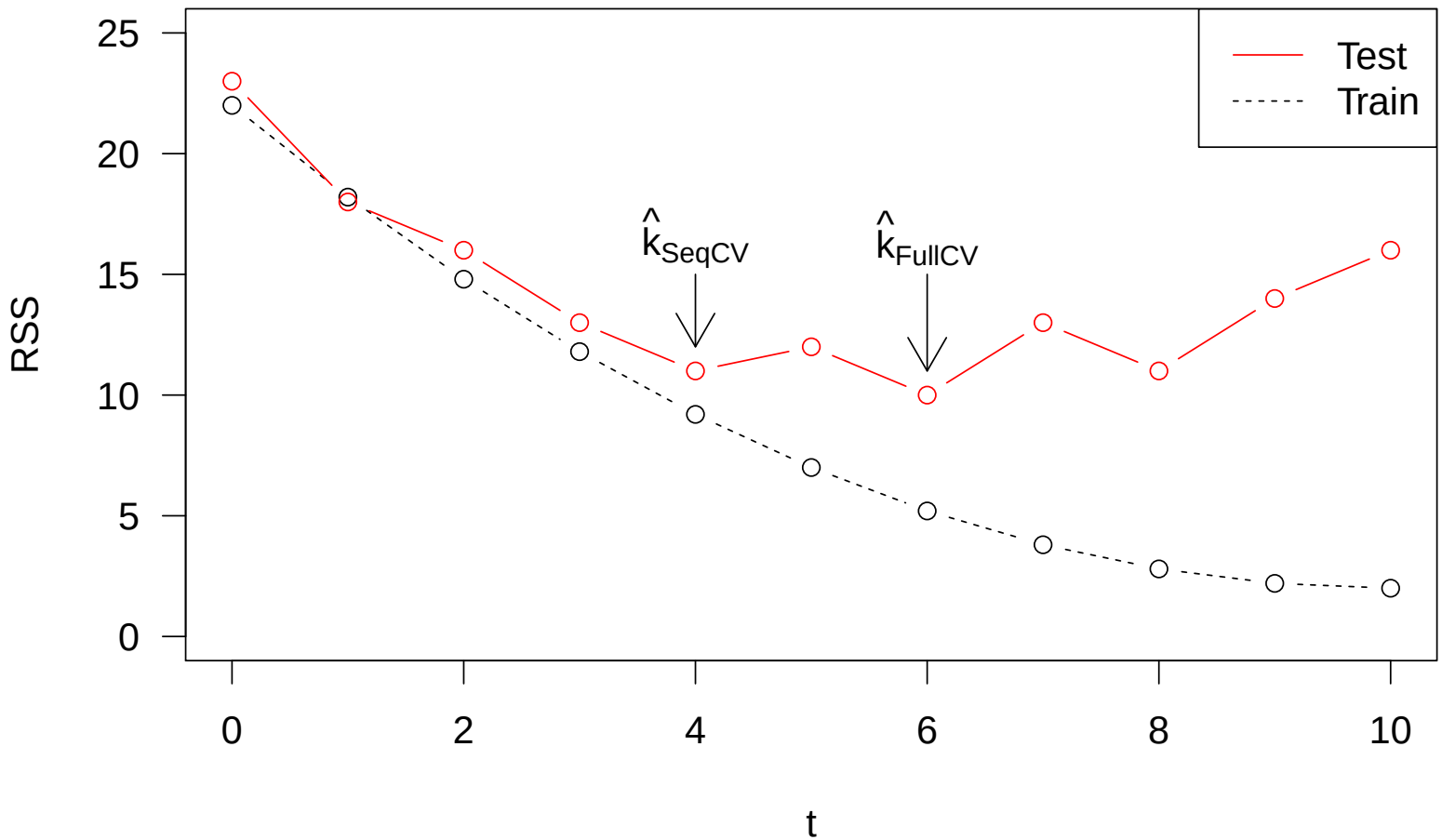
- Randomly split the n cases into *TrainSet* and *TestSet*
- For each model size $t \in 0, \ldots, \min\{n, p\}$:
  - On cases in *TrainSet*, run FS path up to t and fit $\hat{\beta}_{train,t}$
  - Compute $RSS_{test}(t) = \sum_{i \in TestSet}(Y_i - \mathbf{X}_i \hat{\beta}_{train,t})^2$
- Choose model size $\hat{k} = \arg\min_t RSS_{test}(t)$

CV:

- Use some or all of the $\binom{n}{n_{train}}$ possible splits (e.g. V-fold)
- Minimize mean of $RSS_{test}(t, v)$ across splits v

Sequential CV

FS with Sequential CV: instead of $\arg\min_t RSS_{test}(t)$, choose $\hat{k}$ as $\min\{t : RSS_{test}(t) < RSS_{test}(t+1)\}$



Theorem 2: sufficient conditions for FS+SeqCV

Theorem: Assume conditions for FS with known k, and also:

- $\left(\frac{\beta_{min}}{\beta_{max}}\right)^2 \gtrsim \max\left\{k\sqrt{\frac{\log(k)}{n_{train}}}, \frac{k^2\log(k)/n_{test}}{\beta_{min}^2/\sigma^2}\right\}$
- $\frac{p^2}{n_{train}} \cdot k^2\log(p) \to 0$
- $\frac{p^2}{n_{test}/n_{train}} \cdot k\log(p) \to 0$

Then FS with Sequential CV is model-selection consistent.

First condition prevents underfit; next two prevent overfit.
(The disappointing p^2 is sufficient, but unlikely to be necessary. . .)

Theorem 2 proof sketch

Given data $Y, X_1, \ldots, X_p$, compare
true $Y = \beta_0 + \epsilon$ vs. each spurious $Y = \beta_0 + \beta_j X_j + \epsilon$.

We make a mistake if $\exists\ j : \widehat{RSS}_{test}(j) < \widehat{RSS}_{test}(0)$.

Define $\widehat{\Delta}_j = \widehat{RSS}_{test}(j) - \widehat{RSS}_{test}(0)$ and
$\Delta_j = \mathbb{E}_{TestSet}(\widehat{\Delta}_j | TrainSet)$.

If $\frac{p^2}{n_{train}} \asymp 1$, then

$$\mathbb{P}_{TrainSet}\left(\exists\ j :\ \Delta_j < 0\right) \geq 0.12$$

but if $\frac{p^2}{n_{train}} = o(1)$, then

$$\mathbb{P}_{TrainSet}\left(\exists\ j :\ \Delta_j < 0\right) \to 0$$

Theorem 2 proof sketch

Finally, $\Delta_j \approx \frac{1}{\sqrt{n_{train}}}$, with $SE(\widehat{\Delta}_j) \approx \frac{1}{\sqrt{n_{test}}}$.

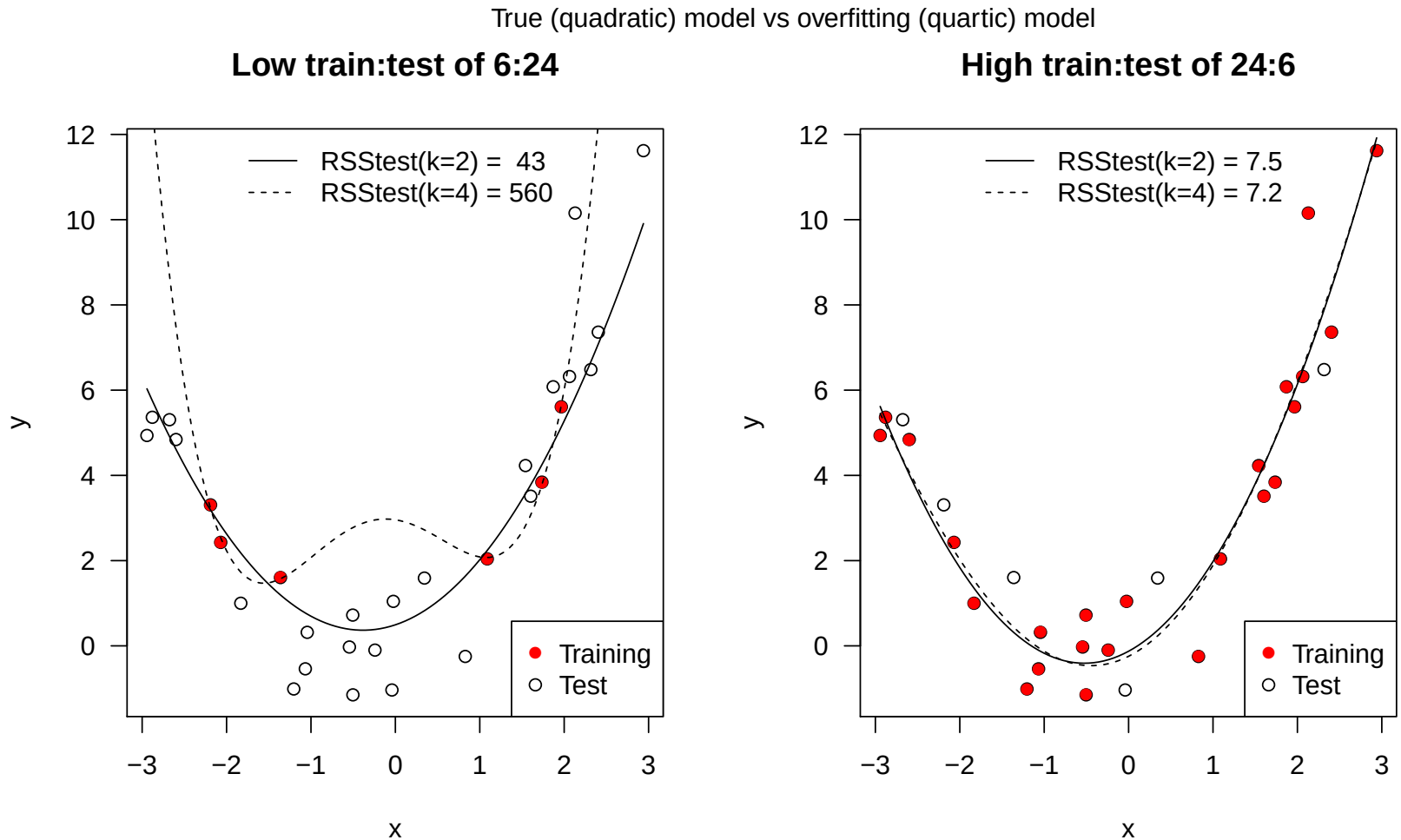
To estimate sign of Δ_j correctly, we need $n_{test} \gg n_{train}$.

By union bound, $\frac{p^2}{n_{test}/n_{train}} = o(1)$ is sufficient for

$$\mathbb{P}\left(\exists\, j : \widehat{\Delta}_j < 0\right) \to 0$$

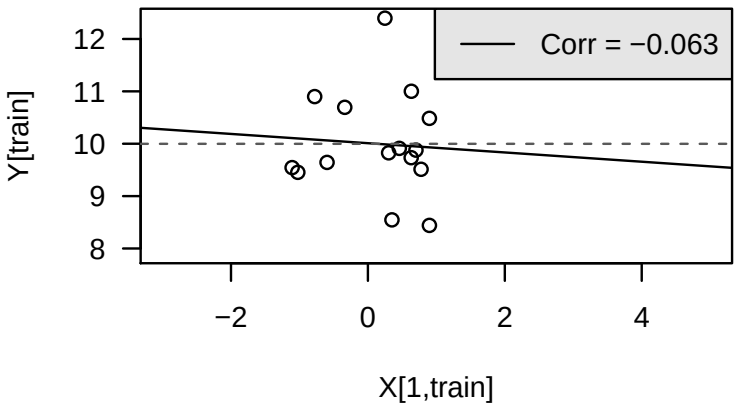
Intuition, illustrated: why $n_{test} \gg n_{train}$?

Compare a correct polynomial in X^2
vs. overfitting polynomial in X^4

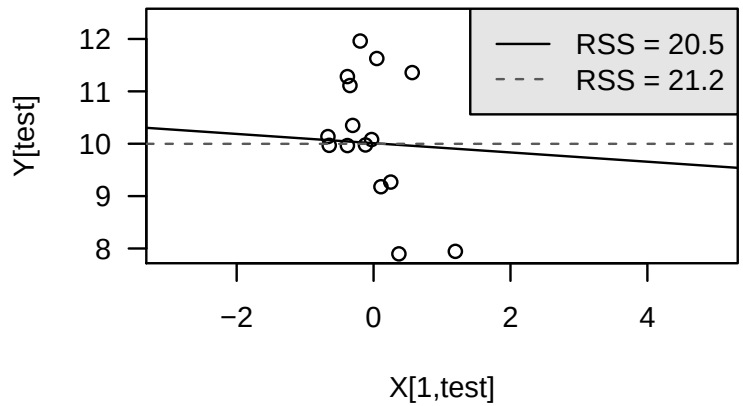


Intuition: why does rising p seem to **help** FS not overfit?

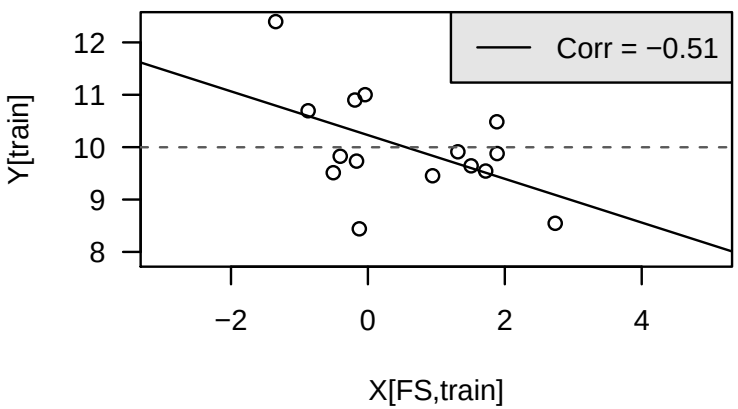
Best fit with low p is likely shallow



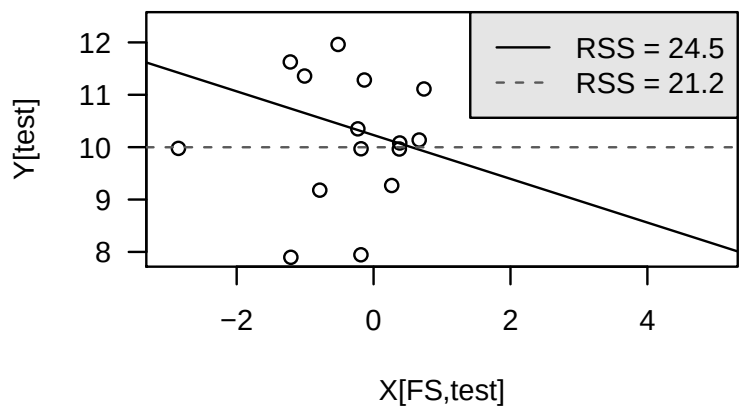
Test data may accept a shallow slope



Best fit with high p may be steep



Test data will reject a steep slope



Summary of theorems

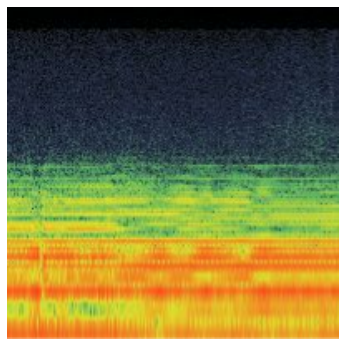
- Is Forward Selection (FS) model-selection consistent?
  - Yes, if model size is known, signal is not too small, and predictors are not too correlated or too numerous
- Is there a consistent stopping rule with few assumptions?
  - Sequential cross-validation (CV) can work if the train:test ratio decreases as sample size grows

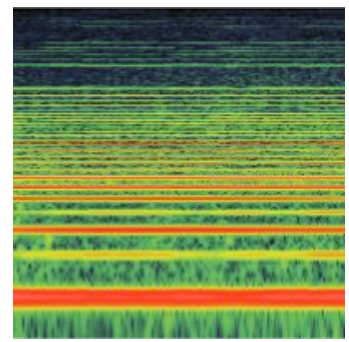
FS+SeqCV is safe if **your predictors are almost uncorrelated** and your sample is big enough to use **a small train:test ratio** (more like 10:90, instead of traditional 90:10).

Traditional high train:test ratios will likely overfit.

Million Song Dataset (Bertin-Mahieux et al., 2011)

- $n \approx 500{,}000$ popular songs, released 1922 to 2011
- Each song summarized by $p = 90$ "timbre" features
- Timbre: the quality of a musical sound that distinguishes different types of musical instruments
- Derived from spectro-temporal surface:
  x = time, y = frequency, color = amplitude



How is timbre associated with the song's release year?
Can we find a good sparse sub-model of the full linear model?

Rule of thumb for choosing training ratio

High training ratios (conventional $\frac{n_{train}}{n_{test}} = 80{:}20$ or $90{:}10$):
lowest Prob(underfit), but non-vanishing Prob(overfit).

Low training ratios ($\frac{n_{train}}{n_{test}} \approx 10{:}90$):
negligible Prob(overfit), but may underfit unless signal is strong, e.g. $\frac{|\beta_{min}|}{\sqrt{k}\cdot\sigma/\sqrt{n}} \geq \sqrt{1 + \frac{n}{n_{train}}}$ for near-orthogonal $\mathbf{X}$.

Rule of thumb:

- Low ratio $\frac{n_{train}}{n_{test}} = 10{:}90$ safe if n large, $\mathbf{X}$ near-orthogonal, and we are confident that $\frac{|\beta_{min}|}{\sqrt{k}\cdot\sigma/\sqrt{n}} \geq \sqrt{1 + \frac{10}{1}} \approx 3.3$
- Else, conventional $\frac{n_{train}}{n_{test}} = 80{:}20$ or $90{:}10$ is safer

MSD example: If true $k \leq 30$, conditions for low train:test ratio seem to be met.

Million Song Dataset: algorithm performance

Algorithm	RMSE (years)	$\hat{k}$	Time (sec)
Null model	10.85	0	—
FS+SeqCV, 10:90	9.60	23	1.3
FS+SeqCV, 90:10	9.56	29	3.2
FS+FullCV, 10:90	9.52	60	3.7
FS+FullCV, 90:10	9.51	76	4.5
Full model	9.51	90	—

SeqCV is sparser and faster than FullCV,
and low training ratio is sparser and faster than high ratio,
with negligible differences in RMSE.

$R^2 \approx 0.22$ to 0.23 in each case,
compared to $R^2 \approx 0.31$ for best non-linear models.

Questions?

Thank you!

Contact: jerzy.wieczorek@colby.edu or @civilstat

Research interests:

- Model selection and evaluation:
  high-dimensional sparse regression, cross-validation, sparse PCA
- Survey sampling methodology:
  ranking with uncertainty, cross-validation for survey data
- Applications in poverty and human-rights data:
  small area estimation, poverty targeting, SWB, DataKind, ICC
- Statistics education:
  curriculum, assessment, pedagogy