Lecture 5

DJM

23 October 2018

Model selection and tuning parameters

- Often "model selection" means "choosing a set of predictors"
  - E.g. Lasso performs model selection by setting many $\widehat{\beta} = 0$
- I define "model selection" more broadly
- I mean "making any necessary decisions to arrive at a final model"
- Sometimes this means "choosing predictors""
- It could also mean "selecting a tuning parameter"
- Or "deciding whether to use LASSO or Ridge" (and picking tuning parameters)
- Recall Lecture 2: "A statistical model $\mathcal{P}$ is a collection of probability
- Model selection means "choose $\mathcal{P}$" distributions or densities."

My pet peeve

- Often people talk about "using LASSO" or "using an SVM"
- This isn't quite right.
- LASSO is a regularized procedure that depends on λ
- To "use LASSO", you must pick a particular λ
- Different ways to pick λ (today's topic) produce different final estimators
- Thus we should say "I used LASSO+CV" or "I used Ridge+GCV"
- Probably also indicate "how" (I used the CV minimum.)

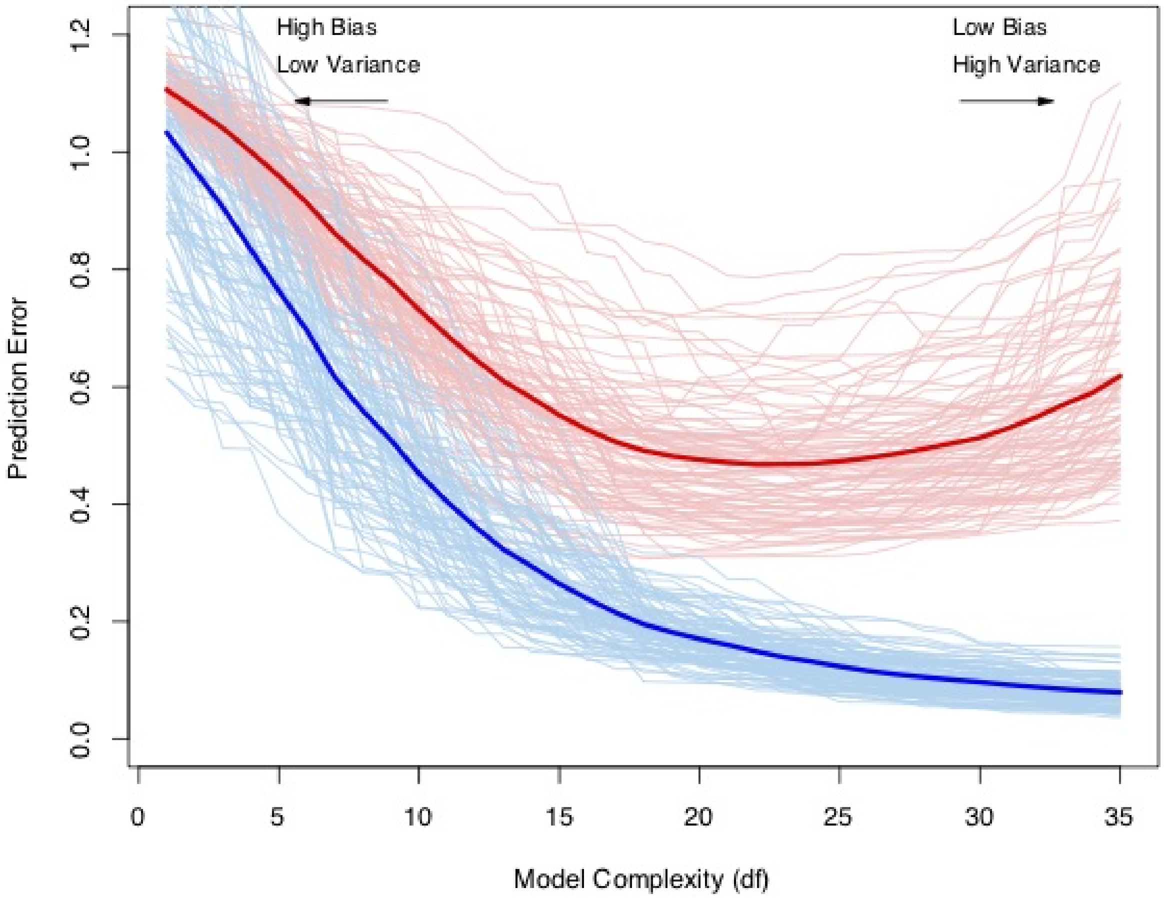
Bias and variance

Recall that $\mathcal{D}$ is the training data.

$$R_n(f) := \mathbb{E}\left[L(Y, f(X))\right] = \mathbb{E}\left[\mathbb{E}\left[L(Y, f(X)) \mid \mathcal{D}\right]\right]$$

- The book calls $R_n(f) = \mathrm{Err}$ and $\mathbb{E}\left[L(Y, f(X)) \mid \mathcal{D}\right] = \mathrm{Err}_{\mathcal{D}}$
- If you use $\mathcal{D}$ to choose f, then these are different.
- If you use $\mathcal{D}$ to choose f, then both depend on how much data you have seen.

Risk estimates

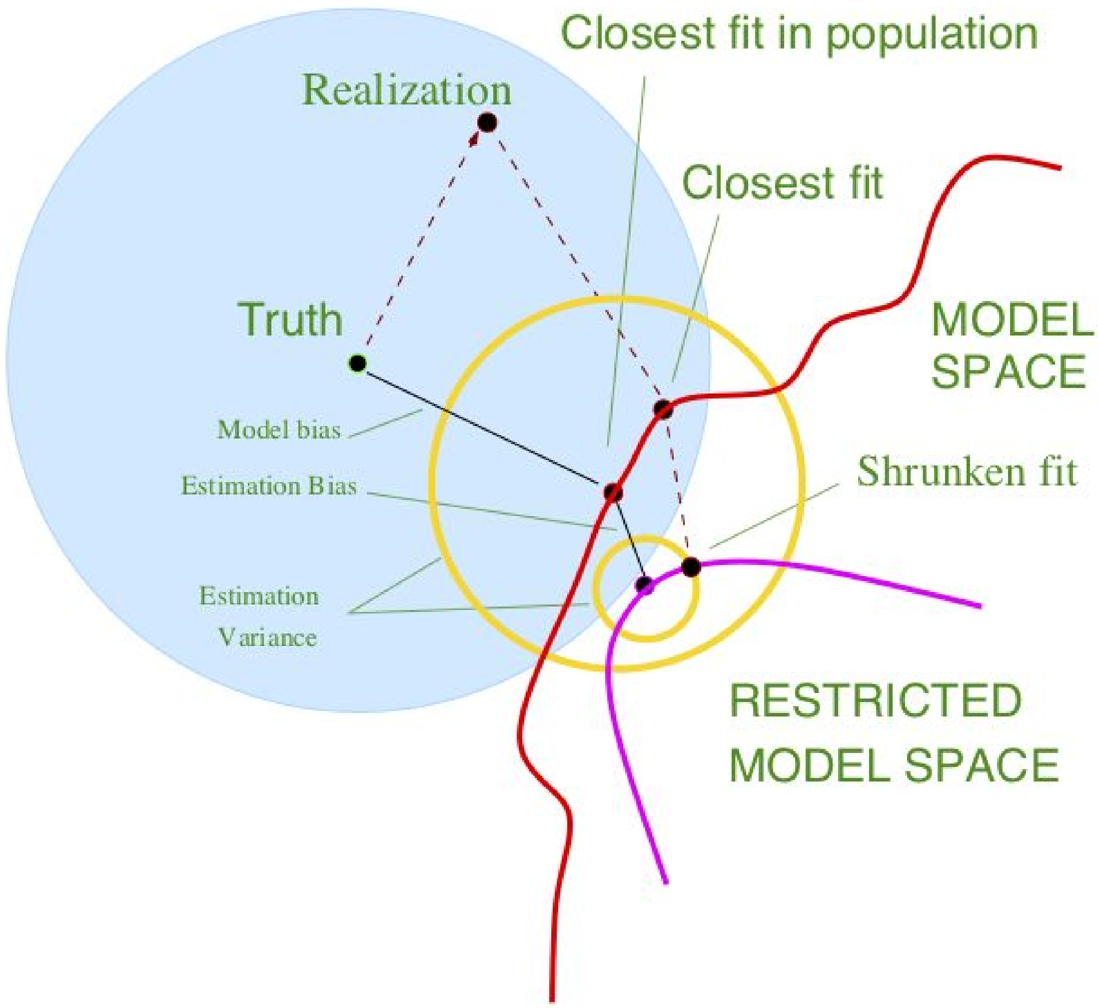


- We can use risk estimates for 2 different goals

1. Choosing between different potential models.
2. Characterizing the out-of-sample performance of the chosen model.

- I am not generally aware of other methods of accomplishing (1).
- You could avoid making a choice (Chapter 8), or you could use a procedure that makes the choice "automatically"
- The method you choose to estimate risk will have large implications for both 1 and 2.

A model selection picture



Why?

We want to do model selection for at least three reasons:

- **Prediction accuracy:** Can essentially *always* be improved by introducing some bias
- **Interpretation:** A large number of features can sometimes be distilled into a smaller number that comprise the "big (little?) picture"
- **Computation:** A large p can create a huge computational bottleneck.

Things you shouldn't do

- Estimate R_n with $\widehat{R}_n(f) = \sum_{i=1}^{n} L(Y_i, \widehat{f}(X_i))$.
- Throw away variables with small p-values.
- Use F-tests
- Compare the log-likelihood between different models

(These last two can occasionally be ok, but aren't in general. You should investigate the assumptions that are implicit in them.)

Risk estimators

Unbiased risk estimation

- It is very hard (impossible?) to estimate R_n.
- Instead we focus on
$$\overline{R}_n(f) = \mathbb{E}_{Y_1,\ldots,Y_n}\left[\mathbb{E}_{Y^0}\left[\frac{1}{n}\sum_{i=1}^n L(Y_i^0, \widehat{f}(x_i)) \mid \mathcal{D}\right]\right].$$
- The difference is that $\overline{R}_n(f)$ averages over the observed x_i rather than taking the expected value over the distribution of X.
- In the "fixed design" setting, these are equal.

For many L and some predictor $\widehat{f}$, one can show

$$\overline{R}_n(\widehat{f}) = \mathbb{E}\left[\widehat{R}_n(\widehat{f})\right] + \frac{2}{n}\sum_{i=1}^n \mathrm{Cov}\left[Y_i,\ \widehat{f}(x_i)\right].$$

This suggests estimating $\overline{R}_n(\widehat{f})$ with
$$\widehat{R}_{gic} := \widehat{R}_n(\widehat{f}) + \mathrm{pen}.$$

If $\mathbb{E}\left[\mathrm{pen}\right] = \frac{2}{n}\sum_{i=1}^n \mathrm{Cov}\left[Y_i,\ \widehat{f}(x_i)\right]$, we have an unbiased estimator of $\overline{R}_n(\widehat{f})$.

Example: Normal means

Normal means model

Suppose we observe the following data:

$$Y_i = \beta_i + \epsilon_i, \qquad i = 1, \ldots, n$$

where $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0,1)$.

We want to estimate
$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n).$$

The usual estimator (MLE) is
$$\widehat{\boldsymbol{\beta}}^{MLE} = (Y_1, \ldots, Y_n).$$

This estimator has lots of nice properties: **consistent, unbiased, UMVUE, (asymptotic) normality...**

MLEs are bad

But, the standard estimator **STINKS!** It's a bad estimator.

It has no bias, but big variance.

$$R_n(\widehat{\boldsymbol{\beta}}^{MLE}) = \text{bias}^2 + \text{var} = 0 + n \cdot 1 = n$$

What if we use a biased estimator?

Consider the following estimator instead:

$$\widehat{\beta}_i^S = \begin{cases} Y_i & i \in S \\ 0 & \text{else.} \end{cases}$$

Here $S \subseteq \{1, \ldots, n\}$.

Biased normal means

What is the risk of this estimator?

$$R_n(\widehat{\boldsymbol{\beta}}^S) = \sum_{i \notin S} \beta_i^2 + |S|.$$

In other words, if some $|\beta_i| < 1$, then don't bother estimating them!

In general, introduced parameters like S will be called **tuning parameters**.

Of course we don't know which $|\beta_i| < 1$.

But we could try to estimate $R_n(\widehat{\boldsymbol{\beta}}^S)$, and choose S to minimize our estimate.

Estimating the risk

By definition, for any estimator $\widehat{\boldsymbol{\beta}}$,

$$R_n(\widehat{\boldsymbol{\beta}}) = \mathbb{E}\left[\sum_{i=1}^n (\widehat{\beta}_i - \beta_i)^2\right]$$

An intuitive estimator of R_n is

$$\widehat{R}_n(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^n (\widehat{\beta}_i - Y_i)^2.$$

This is known as the **training error** and it can be shown that

$$\widehat{R}_n(\widehat{\boldsymbol{\beta}}) \approx R_n(\widehat{\boldsymbol{\beta}}).$$

Also,

$$\widehat{\boldsymbol{\beta}}^{MLE} = \arg\min_{\beta} \widehat{R}_n(\widehat{\boldsymbol{\beta}}^{MLE}).$$

What could possibly go wrong?

Dangers of using the training error

Although

$$\widehat{R}_n(\widehat{\boldsymbol{\beta}}) \approx R_n(\widehat{\boldsymbol{\beta}}),$$

this approximation can be very bad. In fact:

Training Error: $\widehat{R}_n(\widehat{\boldsymbol{\beta}}^{MLE}) = 0$

Risk: $R_n(\widehat{\boldsymbol{\beta}}^{MLE}) = n$

In this case, the **optimism** of the training error is n.

Normal means

What about $\widehat{\boldsymbol{\beta}}^S$?

$$\widehat{R}_n(\widehat{\boldsymbol{\beta}}^S) = \sum_{i=1}^n (\widehat{\beta}_i - Y_i)^2 = \sum_{i \notin S} Y_i^2$$

Well

$$\mathbb{E}\left[\widehat{R}_n(\widehat{\boldsymbol{\beta}}^S)\right] = R_n(\widehat{\boldsymbol{\beta}}^S) - 2|S| + n.$$

So I can choose S by minimizing $\widehat{R}_n(\widehat{\boldsymbol{\beta}}^S) + 2|S|$.

$$\text{Estimate of Risk} = \text{training error} + \text{penalty}.$$

The penalty term corrects for the optimism.

pen in the nice cases

Result:
Suppose $\widehat{f}(x_i) = HY$ for some matrix H, and Y_i's are IID. Then

$$\frac{2}{n}\sum_{i=1}^n \mathrm{Cov}\left[Y_i,\ \widehat{f}(x_i)\right] = \frac{2}{n}\sum_{i=1}^n H_{ii}\mathrm{Cov}\left[Y_i,\ Y_i\right] = \frac{2\mathbb{V}\left[Y\right]}{n}\mathrm{tr}(H).$$

- Such estimators are called "linear smoothers".
- Obvious extension to the heteroskedastic case.
- We call $\frac{1}{\mathbb{V}[Y]}\sum_{i=1}^n \mathrm{Cov}\left[Y_i,\ \widehat{f}(x_i)\right]$ the **degrees of freedom** of $\widehat{f}$.
- Linear smoothers are ubiquitous
- Examples: OLS, ridge regression, KNN, dictionary regression, smoothing splines, kernel regression, etc.

Examples of DF

- OLS

$$H = X^\top (X^\top X)^{-1} X^\top \Rightarrow \mathrm{tr}(H) = \mathrm{rank}(X) = p$$

- Ridge (decompose $X = UDV^\top$)

$$H = X^\top (X^\top X + \lambda I_p)^{-1} X^\top \Rightarrow \mathrm{tr}(H) = \sum_{i=1}^p \frac{d_i^2}{d_i^2 + \lambda} < \min\{p, n\}$$

- KNN df $= n/K$ (each point is it's own nearest neighbor, it gets weight $1/K$)

Finding risk estimators

This isn't the way everyone introduces/conceptualizes prediction risk.

For me, thinking of $\widehat{R}_n$ as overly optimistic and correcting for that optimism is conceptually appealing

We need to also discuss **information criteria**.

In this case one forms a (pseudo)-metric on probability measures.

Comparing probability measures

Kullback-Leibler

Suppose we have data Y that comes from the probability density function f.

What happens if we use the probability density function g instead?

Example:
Suppose $Y \sim N(\mu, \sigma^2) =: f$. We want to predict a new Y_*, but we model it as $Y_* \sim N(\mu_*, \sigma^2) =: g$

How *far* away are we? We can either compare μ to μ_* or Y to Y^*.

Or, we can compute how *far* f is from g.

We need a notion of distance.

Kullback-Leibler

One central idea is **Kullback-Leibler** divergence (or discrepancy)

$$\begin{aligned} KL(f, g) &= \int \log\left(\frac{f(y)}{g(y)}\right) f(y)dy \\ &\propto -\int \log(g(y))f(y)dy \qquad \text{(ignore term without } g) \\ &= -\mathbb{E}_f[\log(g(Y))] \end{aligned}$$

This gives us a sense of the **loss** incurred by using g instead of f.

- KL is not symmetric: $KL(f, g) \neq KL(g, f)$, so it's not a distance, but it is non-negative and satisfies the triangle inequality.

Usually, g will depend on some parameters, call them θ

KL example

- In regression, we can specify $f = N(X^\top \beta_*, \sigma^2)$
- for a fixed (true) β_*,
- let $g_\theta = N(X^\top \beta, \sigma^2)$ over all $\theta = (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}^+$
- $KL(f, g_\theta) = -\mathbb{E}_f[\log(g_\theta(Y))]$, we want to minimize this over θ.
- But f is unknown, so we minimize $-\log(g_\theta(Y))$ instead.

- This is the maximum likelihood value

$$\widehat{\theta}_{ML} = \arg\max_{\theta} g_\theta(Y)$$

- We don't actually need to assume things about a true model nor have it be nested in the alternative models to make this work.

Operationalizing

- Now, to get an operational characterization of the KL divergence at the ML solution

$$-\mathbb{E}_f[\log(g_{\widehat{\theta}_{ML}}(Y))]$$

  we need an approximation (don't know f, still).

Result:
If you maximize the likelihood for a finite dimensional parameter vector θ of length p, then as $n \to \infty$,

$$-\mathbb{E}_f[\log(g_\theta(Y))] = -\log(g_\theta(Y)) + p.$$

- This is AIC (originally "an information criterion", now "Akaike's information criterion").
- Choose the model with smallest AIC
- Often multiplied by 2 "for historical reasons". Ocassionally, given as the negative of this "to be extra annoying".
- Your estimator for θ needs to be the MLE. p includes all estimated parameters.

Back to the OLS example

- Suppose Y comes from the standard normal linear regression model with known variance σ^2.

$$\begin{aligned}
-\log(g_{\widehat{\theta}}) &\propto \log(\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i^\top \widehat{\beta}_{MLE})^2\\
\Rightarrow AIC &= 2\frac{n}{2\sigma^2}\widehat{R}_n + 2p = \widehat{R}_n + \frac{2\sigma^2}{n}p.
\end{aligned}$$

- Suppose Y comes from the standard normal linear regression model with *unknown* variance σ^2. Note that $\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (y_i - x_i^\top \widehat{\beta}_{MLE})^2$.

$$\begin{aligned}
-\log(g_{\widehat{\theta}}) &\propto \frac{n}{2}\log(\widehat{\sigma}^2) + \frac{1}{2\widehat{\sigma^2}}\sum_{i=1}^n (y_i - x_i^\top \widehat{\beta}_{MLE})^2\\
\Rightarrow AIC &\propto 2n\log(\widehat{\sigma}^2)/2 + 2(p+1) \propto \log(\widehat{R}_n) + \frac{2(p+1)}{n}.
\end{aligned}$$

Related quantities

Mallow's Cp

- Defined for linear regression.
- No likelihood assumptions.
- Variance is known

$$C_p = \widehat{R}_n + 2\sigma^2\frac{\mathrm{df}}{n} = AIC$$

Bayes factor

For Bayesian Analysis, we want the posterior. Suppose we have two models A and B.

$$P(B \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid B)P(B)}{P(\mathcal{D})} \propto P(\mathcal{D} \mid B)P(B)$$
$$P(A \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid A)P(A)}{P(\mathcal{D})} \propto P(\mathcal{D} \mid A)P(A)$$

We assume that $P(A) = P(B)$. Then to compare,

$$\frac{P(B \mid \mathcal{D})}{P(A \mid \mathcal{D})} = \frac{P(\mathcal{D} \mid B)}{P(\mathcal{D} \mid A)}.$$

- Called the **Bayes Factor**.
- This is the ratio of marginal likelihoods under the different models.
- Not easy to calculate generally.
- Use the Laplace approximation, some simplifications, assumptions:

$$\log P(\mathcal{D} \mid B) = \log P(\mathcal{D} \mid \widehat{\theta},\ B) - \frac{p\log(n)}{2} + O(1)$$

- Multiply through by -2:

$$BIC = -\log(g_\theta(Y)) + p\log(n) = \log(\widehat{R}_n) + \frac{p\log(n)}{n}$$

- Also called Schwarz IC. Compare to AIC (variance unknown case)

SURE

$$\widehat{R}_{gic} := \widehat{R}_n(\widehat{f}) + \text{pen}.$$

If $\mathbb{E}\left[\text{pen}\right] = \frac{2}{n}\sum_{i=1}^n \text{Cov}\left[Y_i,\ \widehat{f}(x_i)\right]$, we have an unbiased estimator of $\overline{R}_n(\widehat{f})$.

Result: (Stein's Lemma)
Suppose $Y_i \sim N(\mu_i, \sigma^2)$ and suppose f is weakly differentiable. Then

$$\frac{1}{\sigma^2}\sum_{i=1}^n \text{Cov}\left[Y_i,\ \widehat{f}_i(Y)\right] = \mathbb{E}\left[\sum_{i=1}^n \frac{\partial f_i}{\partial y_i}\widehat{f}(Y)\right].$$

- Note: Here I'm writing $\widehat{f}$ as a function of Y rather than x.
- This gives "Stein's Unbiased Risk Estimator"

$$SURE = \widehat{R}_n(\widehat{f}) + 2\sigma^2\sum_{i=1}^n \frac{\partial f_i}{\partial y_i}\widehat{f}(Y) - n\sigma^2.$$

- If $f(Y) = HY$ is linear, we're back to AIC (variance known case)
- If σ^2 is unknown, may not be unbiased anymore. May not care.

CV

What is Cross Validation

- Cross validation
- This is another way or estimating the prediction risk.
- Why?

To recap:

$$R_n(\widehat{f}) = \mathbb{E}[L(Y, \widehat{f}(X))]$$

where the expectation is taken over the new data point (Y, X) and $\mathcal{D}_n$ (everything that is random).

We saw one estimator of R_n:

$$\widehat{R}_n(\widehat{f}) = \sum_{i=1}^{n} L(Y_i, \widehat{f}(X_i)).$$

This is the training error. It is a **BAD** estimator because it is often optimistic.

Intuition for CV

- One reason that $\widehat{R}_n(\widehat{f})$ is bad is that we are using the same data to pick $\widehat{f}$ **AND** to estimate R_n.
- Notice that R_n is an expected value over a **NEW** observation (Y, X).
- We don't have new data.

Wait a minute...

...or do we?

- What if we set aside one observation, say the first one (Y_1, X_1).
- We estimate $\widehat{f}^{(1)}$ without using the first observation.
- Then we test our prediction:

$$\widetilde{R}_1(\widehat{f}^{(1)}) = L(Y_1, \widehat{f}^{(1)}(X_1)).$$

- But that was only one data point (Y_1, X_1). Why stop there?
- Do the same with (Y_2, X_2)! Get an estimate $\widehat{f}^{(2)}$ without using it, then

$$\widetilde{R}_2(\widehat{f}^{(2)}) = L(Y_2, \widehat{f}^{(2)}(X_2)).$$

Keep going

- We can keep doing this until we try it for every data point.
- And then average them! (Averages are good)
- In the end we get

$$\text{LOO-CV} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{R}_i(\widehat{f}^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i - \widehat{f}^{(i)}(X_i))$$

- This is leave-one-out cross validation

Problems with LOO-CV

1. Each held out set is small ($n = 1$). Therefore, the variance of my predictions is high.
2. Since each held out set is small, the training sets overlap. This is bad.
   - Usually, averaging reduces variance:

$$\mathbb{V}\left[\overline{X}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left[X_i\right] = \frac{1}{n}\mathbb{V}\left[X_1\right].$$

   - But only if the variables are independent. If not, then

$$\begin{aligned}\mathbb{V}\left[\overline{X}\right] &= \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\mathbb{V}\left[X_1\right] + \frac{1}{n^2}\sum_{i\neq j}\mathrm{Cov}\left[X_i,\ X_j\right].\end{aligned}$$

   - Since the training sets overlap a lot, that covariance can be pretty big.
3. We have to estimate this model n times.
   - There is an exception to this one. More on that in a minute.

LOO-CV with linear smoothers

Suppose $\widehat{Y} = HY$ and $L(a, b) = (a - b)^2$.

- After much tedious algebra, one can show that

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{R}_i(\widehat{f}^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i - \widehat{f}^{(i)}(X_i)) = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{f}(X_i))^2}{(1 - H_{ii})^2}.$$

- This means we only need to fit the model **once** rather than n times
- This also suggests how to demonstrate that CV and AIC are asymptotically equivalent.
- Suppose $(1 - H_{ii}) \approx (1 - h)\forall i$. Then,

$$\log(\text{LOO-CV}) = \log(\widehat{R}_n) - \log((1 - h)^2)$$

- As n gets large (p-fixed), $h \approx p/n$, and $\log(1 - x) \approx x$ when x is small

$$\Rightarrow -\log((1 - h)^2) \approx 2p/n$$

  just like AIC.

Tedious algebra

Lemma (Sherman-Morrison-Woodbury)

Suppose we have four matrices: A C U V, if A C are invertible and everything conforms, then

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Proof of LOO-CV formula:

$$
\begin{aligned}
y_i - \widehat{y}_{(i)} &= y_i - x_i \widehat{\beta}_{(i)} \\
&= y_i - x_i (X_{(i)}^T X_{(i)})^{-1} x_{(i)}^T y_{(i)} \\
&= y_i - x_i \left((X^T X)^{-1} + \frac{(X^T X) x_i^T x_i (X^T X)^{-1}}{1 - h_{ii}} \right) x_{(i)}^T y_{(i)} \\
&= y_i - x_i \left((X^T X)^{-1} + \frac{(X^T X) x_i^T x_i (X^T X)^{-1}}{1 - h_{ii}} \right) (X^T Y - x_i^T y_i) \\
&= y_i - x_i (X^T X)^{-1} X^T Y + x_i (X^T X)^{-1} x_i^T y_i + \frac{x_i (X^T X) x_i^T x_i (X^T X)^{-1} x_i^T y_i}{1 - h_{ii}} - \frac{x_i (X^T X) x_i^T x_i (X^T X)^{-1} X^T Y}{1 - h_{ii}} \\
&= y_i - x_i \widehat{\beta} + h_{ii} y_i + \frac{h_{ii}^2 y_i}{1 - h_{ii}} - \frac{h_{ii} x_i \widehat{\beta}}{1 - h_{ii}} \\
&= y_i - x_i \widehat{\beta} + \frac{(1 - h_{ii}) h_{ii} y_i}{1 - h_{ii}} + \frac{h_{ii}^2 y_i}{1 - h_{ii}} - \frac{h_{ii} x_i \widehat{\beta}}{1 - h_{ii}} \\
&= y_i - x_i \widehat{\beta} + \frac{h_{ii} y_i}{1 - h_{ii}} - \frac{h_{ii} x_i \widehat{\beta}}{1 - h_{ii}} \\
&= \frac{(y_i - x_i \widehat{\beta})(1 - h_{ii})}{1 - h_{ii}} + \frac{h_{ii} (y_i - x_i \widehat{\beta})}{1 - h_{ii}} \\
&= \frac{y_i - x_i \widehat{\beta}}{1 - h_{ii}} \\
&= \frac{y_i - \widehat{y}_i}{1 - h_{ii}} \\
&= \frac{\widehat{e}_i}{1 - h_{ii}}
\end{aligned}
$$

Generalized Cross-Validation (GCV)

This estimator is close to LOOCV in that we replace $1 - H_{ii}$ by $1 - \frac{1}{n}\mathrm{tr}(H)$ in the equation above.

$$
GCV = \frac{\widehat{R}_n}{\left(1 - \frac{\mathrm{tr}(H)}{n}\right)^2}
$$

- Also asymptotically equivalent to AIC
- Optimal estimator for Ridge regression/RKHS norm regularized smoothers (splines, etc.)
- For selection, tends to dramatically over select.

Generic Cross Validation

Let $\mathcal{N} = \{1, \ldots, n\}$ be the index set for $\mathcal{D}$

Define a distribution $\mathcal{V}$ over $\mathcal{N}$ ($v \sim \mathcal{V} \subseteq \mathcal{N}$)

Then, we can form a general *cross-validation* estimator as

$$
\mathrm{CV}_{\mathcal{V}}(\widehat{f}) = \mathbb{E}\left[\frac{1}{|v|} \sum_{i \in v} L\left(Y_i, \widehat{f}^{(v)}(X_i)\right) \mid \mathcal{V} \right]
$$

More general cross-validation schemes: Examples

$$\mathrm{CV}_{\mathcal{V}}(\widehat{f}) = \mathbb{E}\left[\frac{1}{|v|}\sum_{i\in v} L\left(Y_i, \widehat{f}^{(v)}(X_i)\right) \mid \mathcal{V}\right]$$

- **K-fold:**
  Fix $V = \{v_1, \ldots, v_K\}$ such that $v_j \cap v_k = \emptyset$ and $\bigcup_j v_j = \mathcal{N}$

  $$\mathrm{CV}_K(\widehat{f}) = \frac{1}{K}\sum_{v\in V}\frac{1}{|v|}\sum_{i\in v}(Y_i - \widehat{f}^{(v)}(X_i))^2$$

- **Bootstrap:**
  Let $\mathcal{V}$ be given by the bootstrap distribution over $\mathcal{N}$ (that is, sampling B indices randomly with replacement many times)

- **Factorial:**
  Let $\mathcal{V}$ be given by all subsets (or a subset of all subsets) of $\mathcal{N}$ (that is, putting mass $1/(2^n - 2)$ on each subset)

More general cross-validation schemes: A comparison

- CV_K gets more computationally demanding as $K \to n$
- The bias of CV_K goes down, but the variance increases as $K \to n$
- The factorial version isn't commonly used except when doing a 'real' data example for a methods paper
- There are many other flavors of CV. One of them, called "consistent cross validation" is a recent addition that is designed to work with sparsifying algorithms
- K-fold is most common (like $K = 10$ or $K = 5$)

K-fold CV

1. Divide the data into K groups.
2. Leave a group out and estimate with the rest.
3. Test on the held-out group. Calculate an average risk over these $\sim n/K$ data.
4. Repeat for all K groups.
5. Average the average risks.

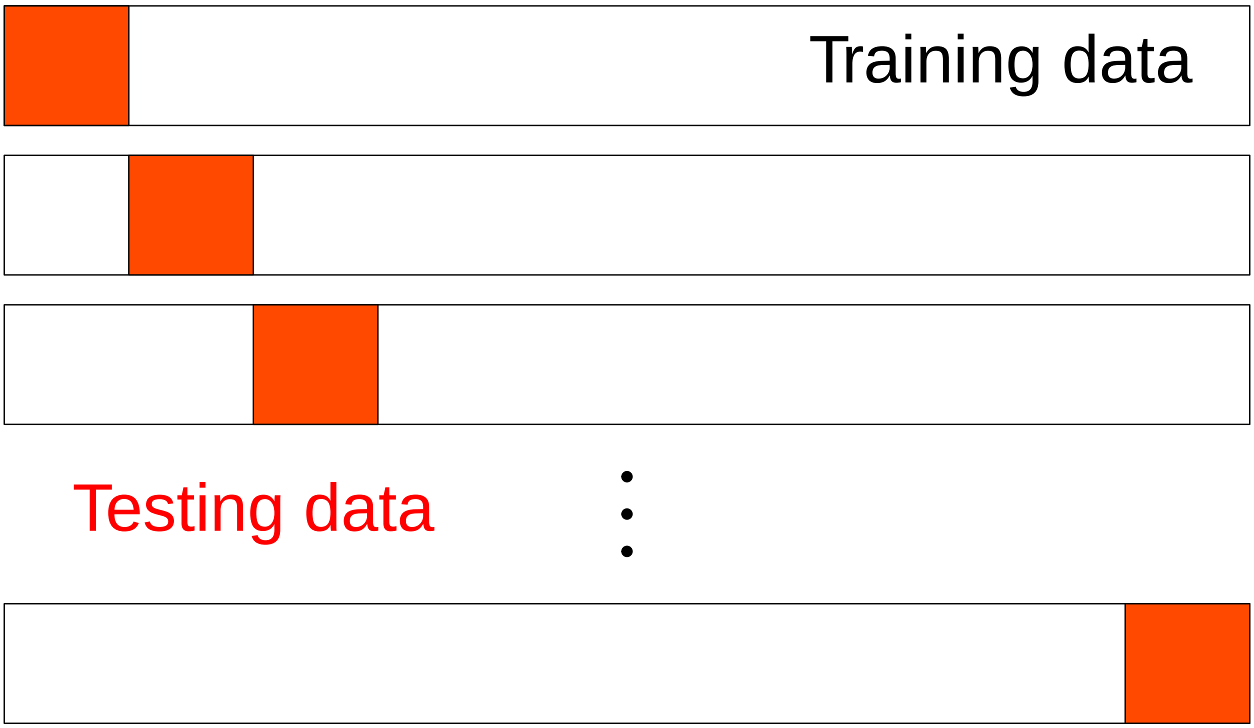
Why K-fold better?

1. Less overlap, smaller covariance.
2. Larger hold-out sets, smaller variance.
3. Less computations (only need to estimate K times)

Why might it be worse?

1. LOO-CV is (nearly) unbiased.
2. The risk depends on how much data you use to estimate the model.
3. LOO-CV uses almost the same amount of data.

A picture



Comparison

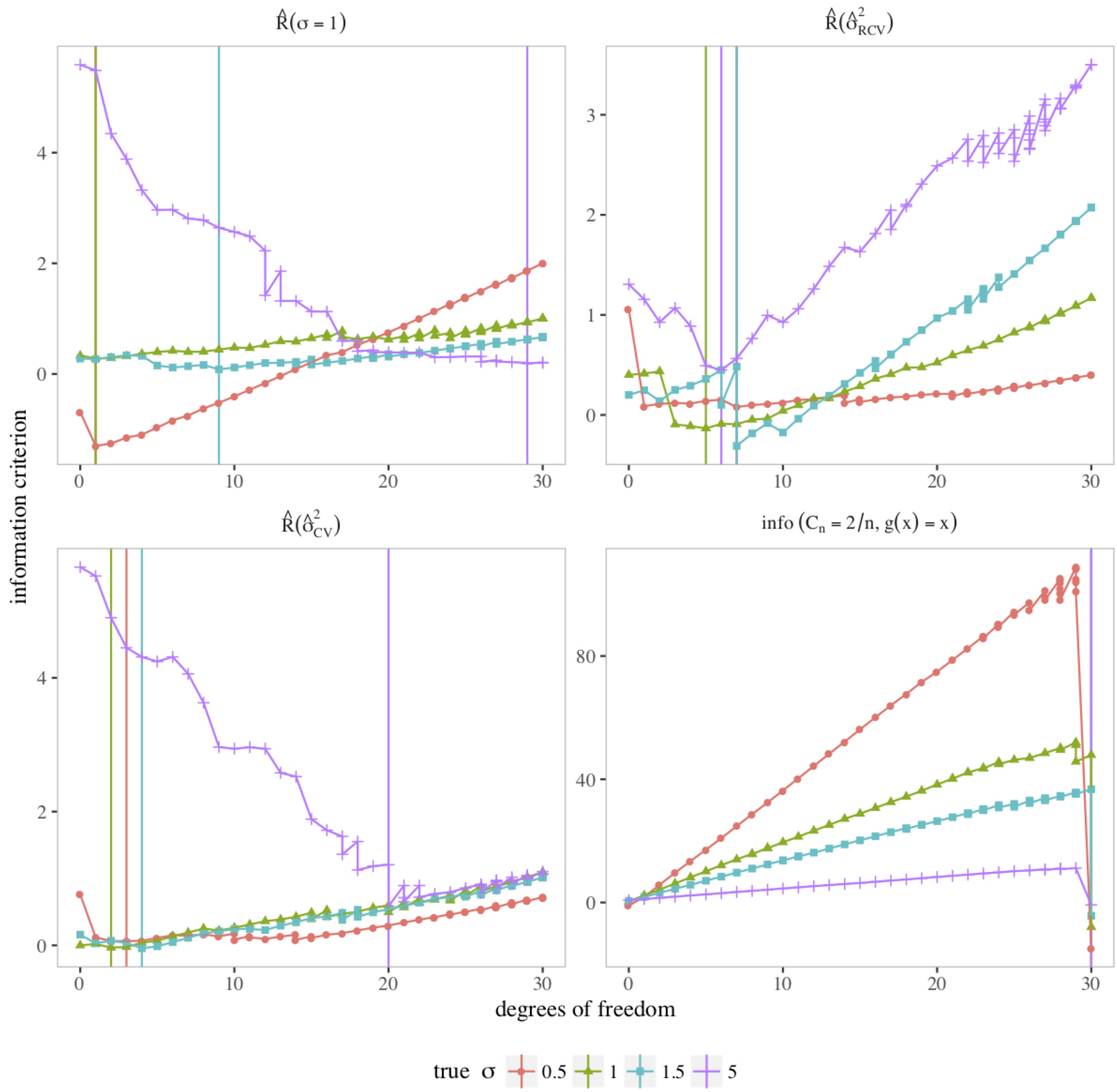
- LOO-CV and AIC are asymptotically equivalent ($n \to \infty$, $p < n$), (Stone 1977)
- Properties of AIC/BIC in high dimensions are not well understood.
- In low dimensions, AIC is minimax optimal for the prediction risk (Yang and Barron 1998)
- CV is consistent for the prediction risk (Dudoit and Laan 2005)
- Both tend to over-select predictors (unproven, except empirically)
- BIC asymptotically selects the correct linear model in low dimensions (Shao 1997) and in high dimensions (Wang, Li, and Leng 2009)
- In linear regression, it is impossible for a model selection criterion to be minimax optimal and select the correct model asymptotically (Yang 2005)
- In high dimensions, if the variance is unknown, the "known" variance form of AIC/BIC is disastrous.
- **Conclusion:** your choice of risk estimator impacts results. Thus,
  1. If you want to select models, you might pick BIC
  2. If you want good predictions, you might use CV
  3. It's possible LASSO+CV(1se) picks models better than LASSO+CV(min)

Some lessons from my work

- The form of AIC I gave you **doesn't** work in high dimensions because you can drive RSS to zero.
- You need to use a high-dimensional variance estimator instead (Homrighausen and McDonald 2018)
- LASSO + CV "works" in high dimensions (not LOO, but no one uses it)
- Under *very* strong conditions it selects the right model at the right rate.
- Under weaker conditions, it achieves (nearly) minimax optimal prediction risk. (Homrighausen and McDonald 2013, 2014, 2017)

AIC/BIC disaster

```
n = 30; p = 150
sigma = c(.5, 1, 1.5, 5)
beta = c(1, 0, ..., 0)
Y = X %*% beta + sigma * rnorm(n)
```



(Brief) foray into model averaging

What if we don't want to choose?

1. Choose a risk estimator $\widehat{R}$

2. Calculate weights $p_i = \exp -\widehat{R}(\text{Model}_i)$
3. Create final estimator $\widehat{f} = \sum_{\text{models}} \frac{p_i}{\sum p_i} \widehat{f}_i$.

- If $\widehat{R}$ is BIC, this is (poor-man's) Bayesian Model Averaging.
- Real BMA integrates over the models:

$$P(f \mid \mathcal{D}) = \int P(f \mid M_i, \mathcal{D}) P(M_i \mid \mathcal{D}) dM$$

- Averaging + Sparsity is pretty hard.
- Interesting open problem: how can we combine LASSO models over the path?
- Issue with MA: e^{-BIC} can be tiny for all but a few models. You're not averaging anymore.

Selected references

Dudoit, Sandrine, and Mark J. van der Laan. 2005. "Asymptotics of Cross-Validation Risk Estimation in Estimator Selection and Performance Assessment." *Statistical Methodology*, 131–54.

Homrighausen, Darren, and Daniel J. McDonald. 2013. "The Lasso, Persistence, and Cross-Validation." In *Proceedings of the* 30^{th} *International Conference on Machine Learning (Icml)*, edited by Sanjoy Dasgupta and David McAllester, 28:1031–9. PMLR.

———. 2014. "Leave-One-Out Cross-Validation Is Risk Consistent for Lasso." *Machine Learning* 97 (1-2): 65–78.

———. 2017. "Risk Consistency of Cross-Validation for Lasso-Type Procedures." *Statistica Sinica* 27 (3): 1017–36.

———. 2018. "A Study on Tuning Parameter Selection for the High-Dimensional Lasso." *Journal of Statistical Computation and Simulation* 88: 2865–92.

Shao, Jun. 1997. "An Asymptotic Theory for Linear Model Selection." *Statistica Sinica* 7 (2): 221–42.

Stone, M. 1977. "Asymptotics for and Against Cross-Validation." *Biometrika* 64 (1): 29–35.

Wang, Hansheng, Bo Li, and Chenlei Leng. 2009. "Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (3): 671–83.

Yang, Yuhong. 2005. "Can the Strengths of Aic and Bic Be Shared? A Conflict Between Model Indentification and Regression Estimation." *Biometrika* 92 (4): 937–50.

Yang, Yuhong, and Andrew Barron. 1998. "An Asymptotic Property of Model Selection Criteria." *IEEE Transactions on Information Theory* 44: 95–116.