Statistical Machine Learning: Introduction

Daniel J. McDonald and Darren Homrighausen

Indiana University, Bloomington
and Colorado State University

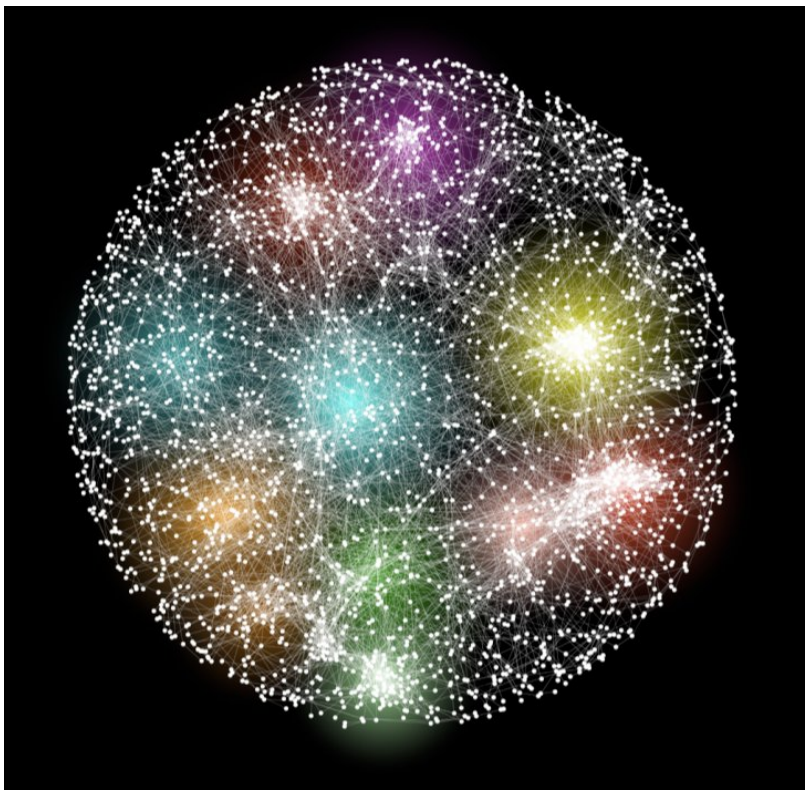
mypage.iu.edu/∼dajmcdon

stat.colostate.edu/∼darrenho

April 2-3, 2013

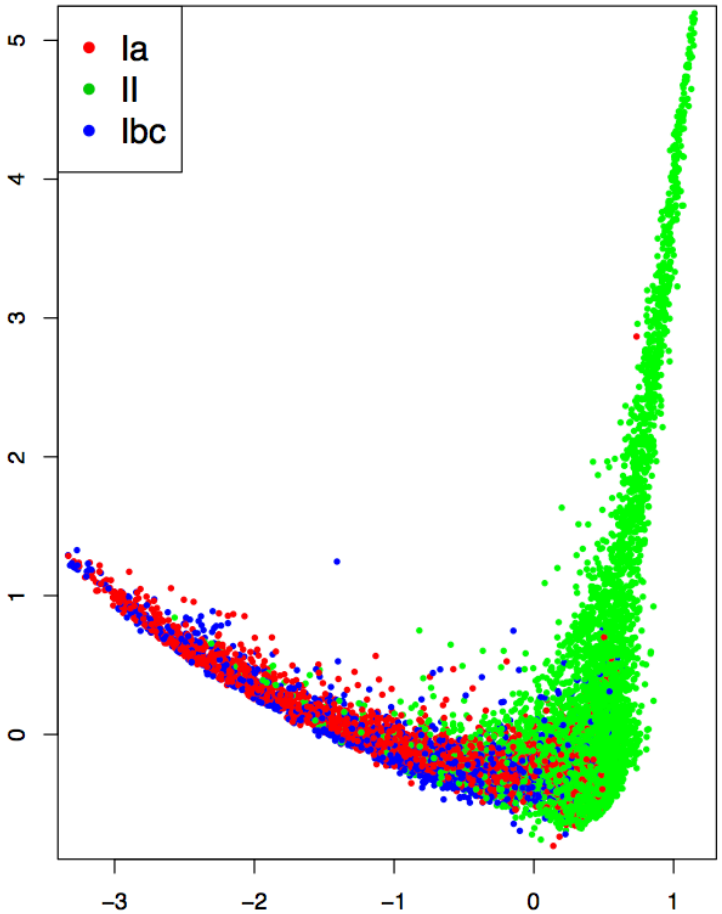
WHAT IS STATISTICS?

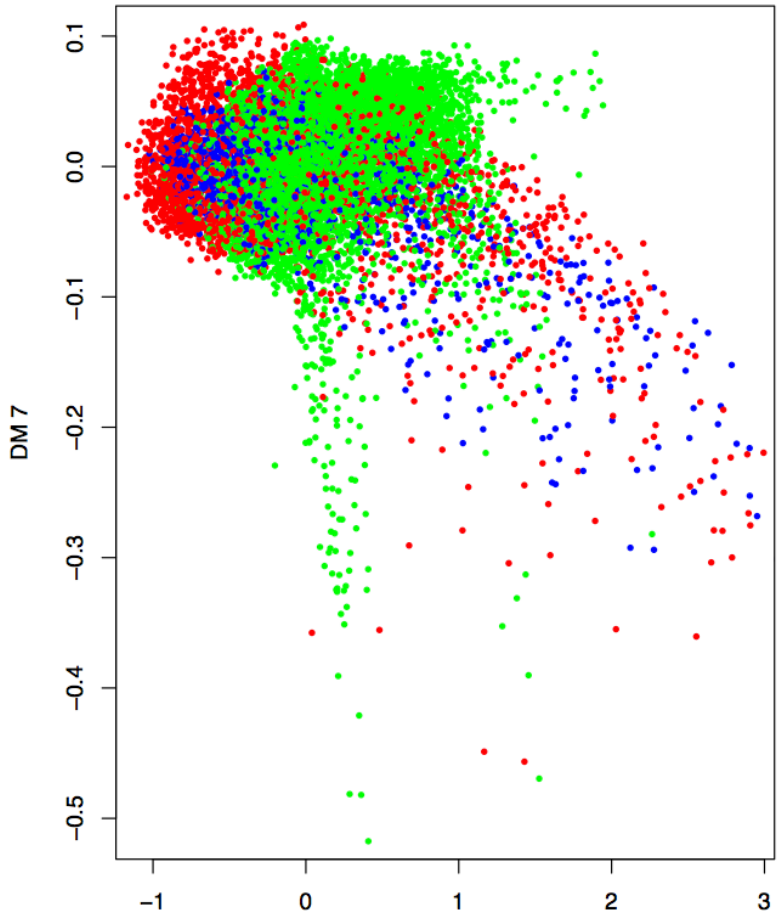
DESCRIPTION Collect some data. Give summaries. Make charts, pretty pictures. Also "unsupervised learning".



WHAT IS STATISTICS?

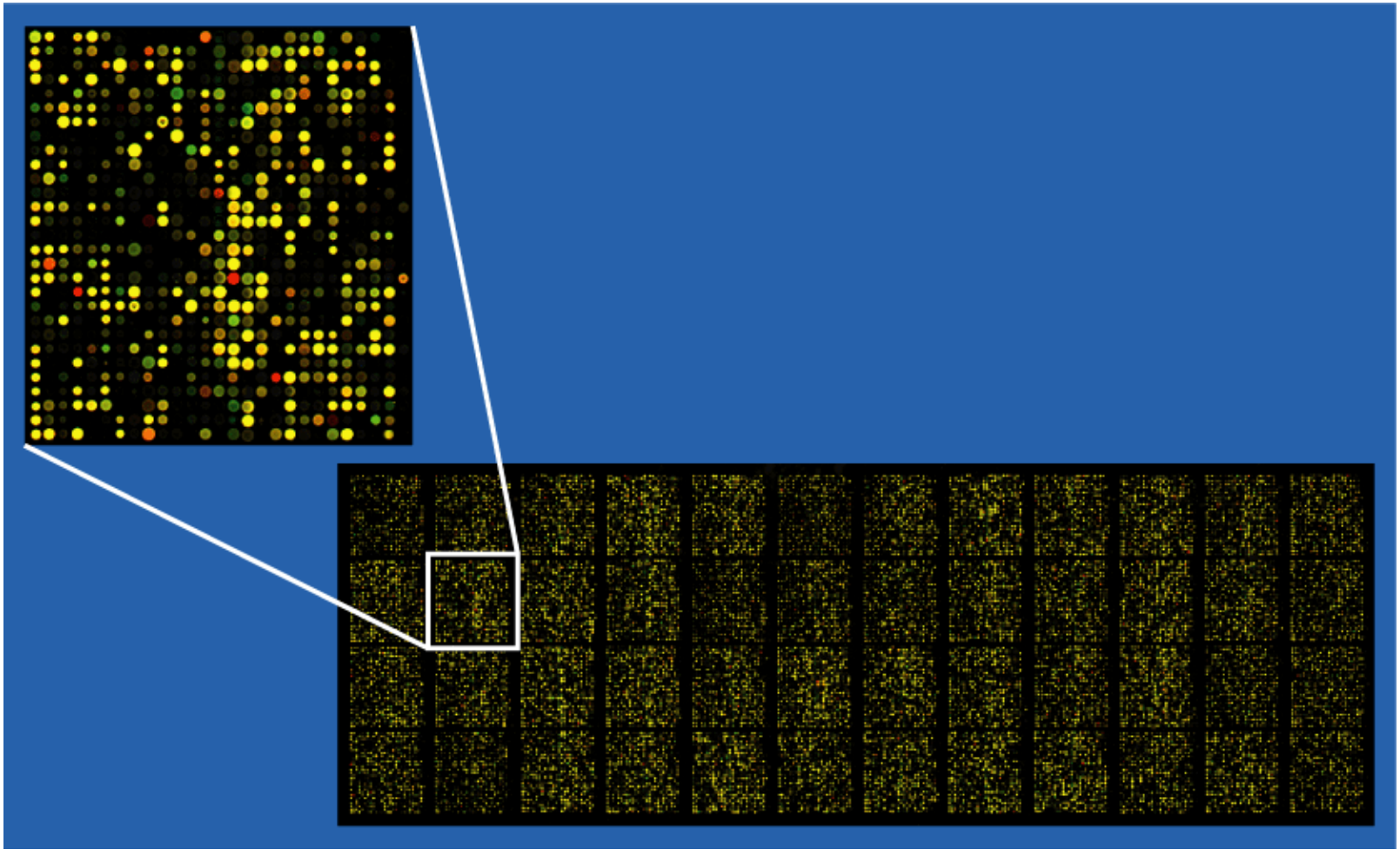
ESTIMATION/INFERENCE Try to determine the underlying causal model.





WHAT IS STATISTICS?

PREDICTION Try to predict some of the data using other data.



What is statistics?

Suppose I come up with some method of description / estimation / prediction. . .

Is it any good?

How can I evaluate my method?

Are there better methods?

Modeling your data

STATISTICAL MODELS

We observe data $Z_1, Z_2, \ldots, Z_n$ generated by some probability distribution P. We want to use the data to learn about P.

A statistical model is a set of distributions $\mathcal{P}$.

Some examples:

1. $\mathcal{P} = \{P(z = 1) = p,\ P(z = 0) = 1 - p,\ 0 < p < 1\}$.
2. $\mathcal{P} = \{Y \sim N(X^\top \beta, \sigma^2),\ \beta \in \mathbb{R}^p,\ \sigma > 0,\ X \text{ fixed}\}$.
3. $\mathcal{P} = \{\text{all CDF's } F\}$.
4. $\mathcal{P} = \{\text{all smooth functions } f : \mathbb{R}^p \to \mathbb{R}\}$

STATISTICAL MODELS

We observe data $Z_1, Z_2, \ldots, Z_n$ generated by some probability distribution P. We want to use the data to learn about P.

$$\mathcal{P} = \{P(z = 1) = p,\ P(z = 0) = 1 - p,\ 0 < p < 1\}$$

To completely characterize P, I just need to estimate p.

Need to assume that $P \in \mathcal{P}$.

This assumption is mostly empty: need independent, can't see $z = 12$.

STATISTICAL MODELS

We observe data $Z_i = (Y_i, X_i)$ generated by some probability distribution P. We want to use the data to learn about P.

$$\mathcal{P} = \{Y \sim N(X^\top \beta, \sigma^2),\ \beta \in \mathbb{R}^p,\ \sigma > 0,\ X \text{ fixed}\}$$

To completely characterize P, I just need to estimate β.

Need to assume that $P \in \mathcal{P}$.

This time, I have to assume a lot more: Linearity, independence, Gaussian noise, no ignored variables, no collinearity, etc

PROPERTIES

Lots of ways to evaluate estimators, $\widehat{\mu}$ of parameters μ.

- Consistency: $\widehat{\mu} \xrightarrow{P} \mu$.
- Asymptotic Normality: $\widehat{\mu} \xrightarrow{D} N(\mu, \Sigma)$
- Efficiency: how large is Σ_μ
- Unbiased: $\mathbb{E}[\widehat{\mu}] \stackrel{?}{=} \mu$
- etc.

None of these things make sense unless your model is correct.

Your model is wrong!

[unless you are flipping coins, gambling in a casino, or running randomized, controlled trials on cereal grains]

MIS-SPECIFIED MODELS

What happens when your model is wrong? And it IS wrong.

None of those evaluation criteria make any sense. The parameters no longer have any meaning.

[The criteria still hold in some sense: I can demand that I get close to the projection of the truth onto $\mathcal{P}$]

PREDICTION

Prediction is easier: your model may not actually represent the true state of nature, but it may still predict well.

Over an 13-year period, [David Leinweber] found, [that annual butter production in Bangladesh] "explained" 75% of the variation in the annual returns of the Standard & Poor's 500-stock index.

By tossing in U.S. cheese production and the total population of sheep in both Bangladesh and the U.S., Mr. Leinweber was able to "predict" past U.S. stock returns with 99% accuracy.

via Carl Richards, NYT 3/26/2012

The predictive viewpoint

THE SETUP

What do we mean by good predictions?

We make observations and then attempt to "predict" new, unobserved data.

Sometimes this is the same as estimating the mean.

Mostly, we observe $(y_1, x_1), \ldots, (y_n, x_n)$, and we want some way to predict Y from X.

EVALUATING PREDICTIONS

Choose some loss function that measures prediction quality:
$\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. We predict Y with $\widehat{Y}$

Examples:

SQUARED-ERROR: $\ell(y, \widehat{y}) = (y - \widehat{y})^2$

ABSOLUTE-ERROR: $\ell(y, \widehat{y}) = |y - \widehat{y}|$

ZERO-ONE: $\ell(y, \widehat{y}) = I(y \neq \widehat{y}) = \begin{cases} 0 & y = \widehat{y} \\ 1 & \text{else} \end{cases}$

Can be generalized to Y in arbitrary spaces.

EVALUATING PREDICTIONS

Of course, both Y and $\widehat{Y}$ are random

I want to know how well I can predict on average

Let $\widehat{f}$ be some way of making predictions $\widehat{Y}$ of Y using covariates X

In fact, suppose I observe a dataset $\mathcal{D}_n = \{(Y_1, X_1,), \ldots, (Y_n, X_n)\}$.
Then I want to choose some $\widehat{f}$ using $\mathcal{D}_n$.

Is $\widehat{f}$ good on average?

Prediction risk

Prediction risk

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

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

For regression applications, we will use squared-error loss:

$$R_n(\widehat{f}) = \mathbb{E}[(Y - \widehat{f}(X))^2]$$

For classification applications, we will use zero-one loss:

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

Example 1: The mean

ESTIMATING THE MEAN

Suppose we know that we want to predict a quantity Y, where $\mathbb{E}[Y] = \mu \in \mathbb{R}$ and $\text{Var}[Y] = \sigma^2 > 0$. That is, $Y \sim P \in \mathcal{P}$, where

$$\mathcal{P} = \{P : \mathbb{E}[Y] = \mu \text{ and } \text{Var}[Y] = \sigma^2\}$$

and σ^2 is known.

Our data is $\mathcal{D}_n = \{Y_1, \ldots, Y_n\}$ such that $Y_i \overset{i.i.d.}{\sim} P$, and we want to estimate μ (and hence P).

Let $\widehat{Y} = \overline{Y}_n$ be the sample mean.

What is the prediction risk?

$$\begin{aligned} R_n(\overline{Y}_n) &= \mathbb{E}[(\overline{Y}_n - Y)^2] = \mathbb{E}[\overline{Y}_n^2] - 2\mathbb{E}[\overline{Y}_n Y] + \mathbb{E}[Y^2] \\ &= \mu^2 + \frac{\sigma^2}{n} - 2\mu^2 + \mu^2 + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right) \end{aligned}$$

ESTIMATING THE MEAN

Suppose we know that we want to predict a quantity Y, where $\mathbb{E}[Y] = \mu \in \mathbb{R}$ and $\text{Var}[Y] = \sigma^2 > 0$. That is, $Y \sim P \in \mathcal{P}$, where

$$\mathcal{P} = \{P : \mathbb{E}[Y] = \mu \text{ and } \text{Var}[Y] = \sigma^2\}$$

and σ^2 is known.

Our data is $\mathcal{D}_n = \{Y_1, \ldots, Y_n\}$ such that $Y_i \overset{i.i.d.}{\sim} P$, and we want to estimate μ (and hence P).

Let $\widehat{Y} = \overline{Y}_n$ be the sample mean.

We can also ask about the estimation risk (since we're estimating μ):

$$\begin{aligned} R_n(\overline{Y}_n; \mu) &= \mathbb{E}[(\overline{Y}_n - \mu)^2] = \mathbb{E}[\overline{Y}_n^2] - 2\mu\mathbb{E}[\overline{Y}_n] + \mu^2 \\ &= \mu^2 + \frac{\sigma^2}{n} - 2\mu^2 + \mu^2 = \frac{\sigma^2}{n} \end{aligned}$$

ESTIMATING THE MEAN

Prediction risk: $R(\overline{Y}_n) = \sigma^2 \left(1 + \frac{1}{n}\right) = \sigma^2 + \frac{\sigma^2}{n}$
Estimation risk: $R(\overline{Y}_n; \mu) = \frac{\sigma^2}{n}$

There is actually a nice interpretation here:

The common σ^2/n term is $\text{Var}[\overline{Y}_n]$

The extra factor of σ^2 in the prediction risk is irreducible error — Y is a random variable, and hence noisy. We can never eliminate it's intrinsic variance. In other words, even if we knew μ, we could never get closer than σ^2, on average.

ESTIMATING THE MEAN

Suppose we consider a different prediction $\widehat{Y}_a = a\overline{Y}_n$ for $a \in (0, 1)$.

$$R_n(\widehat{Y}_a) = \mathbb{E}[(\widehat{Y}_a - Y)^2] = (1-a)^2\mu^2 + \frac{a^2\sigma^2}{n} + \sigma^2$$

We can minimize this in a to get the best possible prediction risk for an estimator of the form $\widehat{Y}_a$:

$$\operatorname*{argmin}_a R_n(\widehat{Y}_a) = \left(\frac{\mu^2}{\mu^2 + \sigma^2/n}\right)\overline{Y}_n$$

What happens if $\mu \ll \sigma$?

Wait a minute! You're saying there is a better estimator than $\overline{Y}_n$?

Of course to compute this estimator, we need to know μ, which is what we are estimating! But we don't know μ... (more on this later).

Why deal with prediction risk?

PREDICTION RISK

Why care about $R_n(f)$?

Measures predictive accuracy on average.

How much confidence should you have in f's predictions.

Compare with other models.

This is hard:
Don't know P (if I knew the truth, this would be easy)

WHAT IF YOU REALLY WANT TO MAKE INFERENCES?

1. You don't really care about predicting what will happen next year / quarter / millisecond
2. But you do want to offer an explanation / evaluate counterfactuals / describe the world
3. So you need the structure of your model to be at least approximately right
4. If you cannot predict well, then your model cannot be correct at all
5. Therefore, a necessary condition to believe your counterfactuals is good predictive accuracy

- Step 4 is about prediction error. (Sum of squared residuals??)
- Best not be fooling yourself in step 5.

RISK FOR GENERAL MODELS

We just saw that when you know the true model, and you have a nice estimator, the prediction risk has a nice decomposition
(this generalizes to much more complicated situations)

- Suppose we have a class of prediction functions $\mathcal{F}$,

$$\text{e.g. } \mathcal{F} = \left\{\widehat{f}(x) = \beta^\top x \text{ for some } \beta\right\}$$

- We use the data to choose some $\widehat{f} \in \mathcal{F}$ and set $\widehat{Y} = \widehat{f}(X)$
- The true model is g (not necessarily in $\mathcal{F}$). Then:

$$R_n(\widehat{f}) = \int \left[\text{bias}^2(\widehat{f}(x)) + \text{var}(\widehat{f}(x))\right] p(x)dx + \sigma^2$$

where $X \sim p$ and

$$\begin{aligned}
\text{bias}(\widehat{f}(x)) &= \mathbb{E}[\widehat{f}(x)] - g(x) \\
\text{var}(\widehat{f}(x)) &= \mathbb{E}[(\widehat{f}(x) - \mathbb{E}\widehat{f}(x))^2] \\
\sigma^2 &= \mathbb{E}[(Y - g(X))^2]
\end{aligned}$$

BIAS-VARIANCE DECOMPOSITION

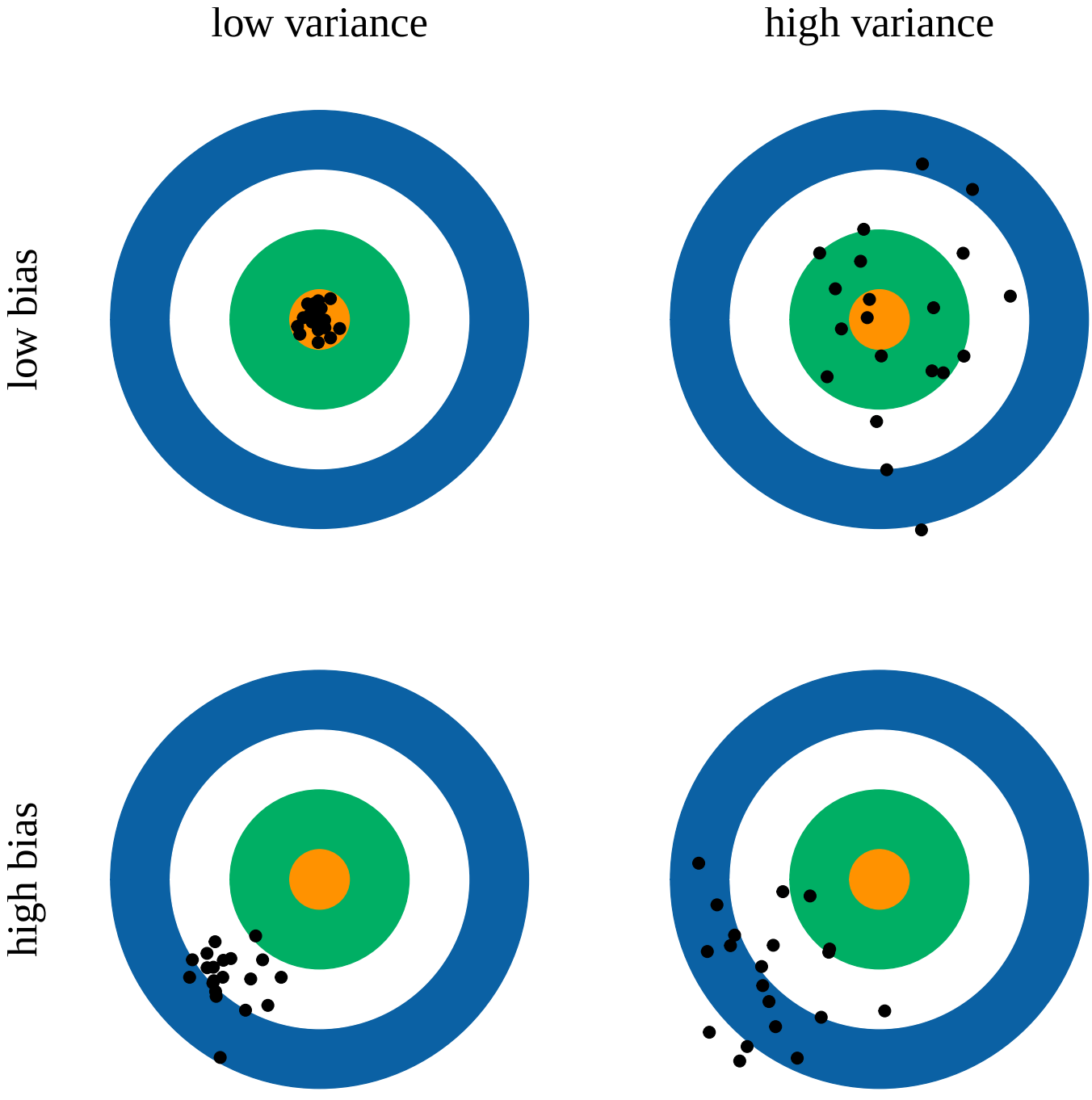
So,

$$\begin{array}{lcllcl} \text{prediction risk} & = & \text{bias}^2 & + & \text{variance} & + \quad \text{irreducible error} \\ \text{estimation risk} & = & \text{bias}^2 & + & \text{variance} & \end{array}$$

$$\left(\text{Compare to: } R_n(\widehat{Y}_a) = (1-a)^2\mu^2 + \frac{a^2\sigma^2}{n} + \sigma^2 \right)$$

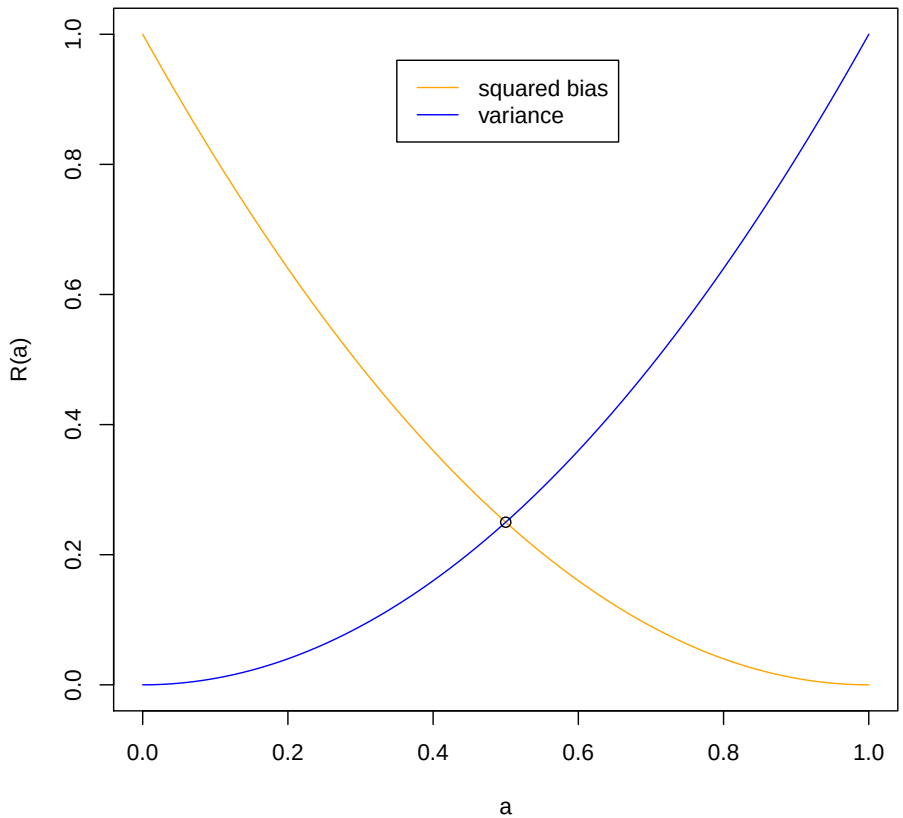
Important implication: prediction risk is proportional to estimation risk. However, defining estimation risk requires stronger assumptions.

In order to make good predictions, we want our prediction risk to be small. This means that we want to 'balance' the bias and variance.


low variance
high variance
low bias
high bias

BIAS-VARIANCE TRADEOFF: ESTIMATING THE MEAN

$$R(a) = R_n(\widehat{Y}_a) = (1-a)^2\mu^2 + \frac{a^2\sigma^2}{n} + \sigma^2$$



Here, $\mu = \sigma^2 = 1$, and hence optimal $a = .5$. This corresponds to the intersection of these curves.

BIAS-VARIANCE TRADEOFF: OVERVIEW

- bias: how well does $\widehat{f}$ approximate the truth g
- more complicated $\mathcal{F}$, lower bias. Flexibility $\Rightarrow$ Parsimony
- more flexibility $\Rightarrow$ larger variance
- complicated models are hard to estimate precisely for fixed n
- irreducible error

Example 2: Normal means

NORMAL MEANS

Suppose we observe the following data:

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

where $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ and σ^2 is known.

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

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

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

But the MLE STINKS! It's a bad estimator.

It has no bias, but big variance.

$$R_n(\widehat{\beta}^{MLE}) = \text{bias}^2 + \text{var} = 0 + n\sigma^2 = n\sigma^2.$$

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

NORMAL MEANS

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

In other words, if some $\beta_i < \sigma^2$, then don't bother estimating them!

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

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

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

ESTIMATING R_n

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,

$$\widehat{\boldsymbol{\beta}}^{MLE} = \underset{\beta}{\operatorname{argmin}}\, \widehat{R}_n(\widehat{\boldsymbol{\beta}}^{MLE}).$$

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:

$$\begin{array}{lrcl} \text{TRAINING ERROR:} & \widehat{R}_n(\widehat{\boldsymbol{\beta}}^{MLE}) & = & 0 \\ \text{RISK:} & R_n(\widehat{\boldsymbol{\beta}}^{MLE}) & = & n\sigma^2 \end{array}$$

In this case, the optimism of the training error is $n\sigma^2$.

NORMAL MEANS

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

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

Well

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

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

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

The penalty term corrects for the optimism.

Where we're going

Themes of course

Bias is good

1. Very often, we can trade some bias for (much) lower variance
2. Bias is controlled by setting tuning parameters (e.g. S)
3. Choosing tuning parameters carefully gives good risk properties
4. To know how to choose the tuning parameters, we need an estimate of the risk
5. Training error (which a risk estimator) is a bad choice (optimistic)
6. Unbiased estimators of parameters in correct models may have nice properties
7. Unbiased estimators of parameters in mis-specified models rarely have nice properties
8. All models are mis-specified

Up next...

1. Regression and regularization
2. Economic forecasting and time series
3. Classification
4. More fun stuff...