Risk bounds for time series

Daniel J. McDonald

Department of Statistics
Indiana University, Bloomington
http://mypage.iu.edu/∼dajmcdon

Joint work with:
Cosma Shalizi and Mark Schervish

October 9, 2012

What is statistics?

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

Estimation/Inference Try to determine the underlying causal model.

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

Statistical models

We observe data $Z_1, Z_2, \ldots, Z_n \sim F$. We want to use the data to learn about F.

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

Some examples:

1. $\mathcal{F} = \left\{ f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \ \ \mu \in \mathbb{R}, \ \sigma > 0 \right\}$.

2. $\mathcal{F} = \{ Y \sim N(X\beta, \sigma^2), \ \beta \in \mathbb{R}^p, \ \sigma > 0 \}$.

3. $\mathcal{F} = \{\text{all CDF's } F\}$.

Parametric estimation

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

If I want to estimate β_3 well, I need lots of assumptions.

[[No ignored variables, no collinearity, linearity, independence, etc.]]

If everything is nice, can show that with probability at least $1 - \delta$,

$$\left|\widehat{\beta}_3 - \beta_3\right| \leq C_\delta \sqrt{\frac{1}{n}}.$$

Where $\widehat{\beta}_3$ is the ordinary least squares estimator (MLE)

Nonparametric estimation

$$\mathcal{F} = \{\text{all CDF's } F\}$$

If I want to estimate the CDF well, I don't need many assumptions

[[just IID data]]

Can show that with probability at least $1 - \delta$,

$$\sup_x \left| \widehat{F}(x) - F(x) \right| \leq C_\delta \sqrt{\frac{1}{n}}.$$

Where $\widehat{F}$ is the empirical CDF.

Rates

In estimation problems, there are many properties that people look at

- Consistency
- Asymptotic Normality
- Efficiency
- Unbiased
- etc.

Most of these come for free in parametric problems where the model is correct.

Rates

Rates are often more interesting

For parametric problems, one can show that the rate that your estimator gets close to the true value is

$$O_P\left(\frac{1}{\sqrt{n}}\right).$$

In fact this is the best possible.

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 $\mathcal{F}$]]

Prediction

Prediction is easier: your model can be garbage, 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 setup

What do I mean by good predictions?

I get data, and I want to "predict" some function of the CDF F.

I could predict the mean $\mu(F)$. 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.

Good predictions

Choose some model $\mathcal{F}$.

I get to pick a particular $f \in \mathcal{F}$, and I want to know how well that f will predict, given a new pair (Y, X) from the same distribution that generated the observed data.

Prediction Risk

The prediction risk of a function f for predicting Y from X, with loss ℓ and data-source $\mathbb{P}$ is

$$R_n(f) := \mathbb{E}_{\mathbb{P}}\left[\ell(Y, f(X))\right].$$

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 $\mathbb{P}$ (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. The fit between your model and the data is so compelling you'd have to be crazy to think it didn't get the structure at least approximately right
5. And therefore I should believe your counterfactuals

This is all about not fooling yourself in step (4)

Empirical risk

How do I find $R_n(f)$?

Could estimate it with

$$\widehat{R}_n(f) := \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

If f is fixed beforehand, then

$$|R_n(f) - \widehat{R}_n(f)| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Parametric!

Empirical risk

I just showed that for a fixed function f,
I can estimate the risk $R_n(f)$ well, despite knowing nothing about $\mathbb{P}$.

But that doesn't mean that $R_n(f)$ will be small.

How do we make $R_n(f)$ small?

Bound the risk

For fixed f,

$$|R_n(f) - \widehat{R}_n(f)| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

This means, with probability at least $1 - \delta$,

$$R_n(f) \leq \widehat{R}_n(f) + C_\delta \sqrt{\frac{1}{n}}$$

I need both $\widehat{R}_n(f)$ and $C_\delta \sqrt{\frac{1}{n}}$ to be small, then $R_n(f)$ must be small with high probability.

I just need better $f \in \mathcal{F}$ to drive $\widehat{R}_n(f) \searrow 0$.

Bias and variance

In the case above $\mathcal{F} = \{f\}$ is really small, so I get parametric rates.

If $\mathcal{F}$ gets big, and I get to choose $\widehat{f}$ using the observed data, say by

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i),$$

[[or Bayesian methods, or maximum likelihood, or haruspicy...]]

Then I don't get that nice $C_\delta \sqrt{\frac{1}{n}}$ term anymore.

TRADEOFF

This is the well-known "bias-variance tradeoff".

By making $\mathcal{F}$ large, I lower the risk of the best possible estimator.

By making $\mathcal{F}$ large, I increase the variance of my chosen estimator, $\widehat{f}$. There are too many functions in $\mathcal{F}$, so I end up picking the one that fits the observed data most closely. I reproduce all the noise in the sample.

Suppose $\ell(y', y) = (y - y')^2$, $Y = f(X) + \epsilon$ (this is always true). Then,

$$R_n(f) = \text{irreducible noise} + \text{approx. error of } \mathcal{F} + \\ + \text{estimation bias}^2 + \text{estimation variance}.$$

Utility

So how is all this useful/interesting for economics?

One way:

Economists try to predict economic activity.

They use DSGEs.

These DSGEs fail to predict the financial collapse.

[[So do VARs and Dynamic Factor models and whatever else.]]

Robert Solow and others testify before congress.

"Our models are too simple. More complicated models will be better."

Economists build "financial sector" into DSGEs

Good or Bad

So how do I know if my new and improved DSGE has higher or lower prediction risk than my old DSGE?

Are either better than a VAR?

Am I better off just using the long-run mean?

Literature sketch

AIC, BIC, Bayes Factors, FPE, etc. — These are asymptotic. Don't tell you anything about $n \approx 250$. Assume that $\mathcal{F}$ is well specified.
Hold out sets — Not really "held out" in economics. Plus $n \approx 250$. (For comparison, Netflix prize used $n = 10^8$ and a held out set of 3 million).
Risk bounds — Hold for finite n. Hold uniformly over $\mathbb{P}$. Don't assume $\mathcal{F}$ is well-specified.
Risk bounds — In literature for weird models and IID data
This stuff works — How does Google choose amongst different models?

Plan of talk

1. Building risk bounds
2. Handling dependence
3. Measuring capacity
4. Our central result
5. Extensions and other considerations
6. Summary

Statistical Learning Theory

THE BASIC FORM OF STATISTICAL LEARNING THEORY

Get data $(x_1, y_1), \ldots, (x_n, y_n)$. Choose class of functions $\mathcal{F}$.
Empirical risk of a fixed function (not data dependent):

$$\widehat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) = R(f) + \gamma_n(f)$$

$$\gamma_n(f) := \text{mean zero idiosyncratic noise}$$

Deviation inequalities for fixed functions:

$$\mathbb{P}\left(|\widehat{R}_n(f) - R(f)| > \epsilon \right) \leq e^{-g(n, \mathcal{F}, \epsilon)}$$

Typically $g(n, \mathcal{F}, \epsilon) = Cn\epsilon^2$.

Union bounds

All well and good, but what about functions chosen using the data?
Often select:

$$\widehat{f} := \underset{f \in \mathcal{F}}{\operatorname{argmin}} \widehat{R}_n(f) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ R(f) + \gamma_n(f) \right\}$$

Suppose:

1. $|\mathcal{F}|$ is finite.
2. For each $f \in \mathcal{F}$,

$$\mathbb{P}\left(|\widehat{R}_n(f) - R(f)| > \epsilon \right) \leq e^{-g(n,\mathcal{F},\epsilon)}.$$

Then, apply union bound to get

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| > \epsilon \right) \leq |\mathcal{F}| e^{-g(n,\mathcal{F},\epsilon)}.$$

$|\mathcal{F}|$ NOT FINITE

Limited capacity: number of effectively distinct f in $\mathcal{F}$ is small
Could even grow (slowly) with n, call this number $G(n, \mathcal{F})$
Then,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |R(f) - \widehat{R}_n(f)| > \epsilon\right) \leq G(n, \mathcal{F})e^{-g(n,\mathcal{F},\epsilon)}$$

Trade off precision [depends on ϵ] and confidence [depends on n, ϵ]

Invert to get confidence bounds

Typically: with probability at least $1 - \eta$,

$$R(\widehat{f}) \leq \widehat{R}_n(\widehat{f}) + C\sqrt{\frac{\log G(n, \mathcal{F}) + \log 1/\eta}{n}}$$

If you insist on asymptotics...

Uniform LLN:

$$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| \to 0$$

Risk-consistency:

$$\text{optimal risk } R^* := \inf_{f \in \mathcal{F}} R(f)$$

and so

$$\left| \widehat{R}_n(\widehat{f}) - R^* \right| \to 0$$

What do we need to make this work?

1. A pointwise deviation inequality (finite-sample law of large numbers)
   Holds for each $f \in \mathcal{F}$
2. A way of saying how big the model $\mathcal{F}$ is

These are extensively developed for IID data and for CS-style models

support vector machines, etc.

We need to handle dependent data and the usual sort of time-series models

Handling dependence

Breeding dependent LLNs

Key assumption: data come from a stationary β-mixing (absolutely regular) process

$$\beta_a = \left\| \mathbb{P}_{-\infty:0 \otimes a:\infty} - \mathbb{P}_{-\infty:0} \times \mathbb{P}_{a:\infty} \right\|_{TV},$$

[[Introduced in 1950s to study central limit theorem etc. for dependent data]]

β-mixing process: $\beta_a \to 0$ as $a \to \infty$

Intuition: at large separations, events are nearly independent

THE BLOCKING TRICK (PROOF TECHNIQUE ONLY)

1. Divide $(Y_1, Y_2, \ldots Y_n)$ into 2μ blocks of length a

   [[Choose μ, a s.t. $2\mu a \leq n$]]

2. Dependence between blocks $\leq \beta_a$
3. Approximate probabilities of events Z over dependent blocks, $\mathbb{P}(Z)$ with probabilities over IID blocks, $\widetilde{\mathbb{P}}(Z)$
   Then by a nice theorem,[1]

$$|\mathbb{P}(Z) - \widetilde{\mathbb{P}}(Z)| \leq \beta_a \mu$$

Intuition: n mixing samples $\approx \mu < n$ independent samples

$\therefore$ we can use IID laws with small corrections

[1] YU (1994), *Rates of Convergence for Empirical Processes of Stationary Mixing Sequences*

Where do the mixing coefficients come from?

- In this talk, assume β_a is given
- Mixing is known for models like ARMA, linear-Gaussian state space models, GARCH, stochastic volatility, ...
- Could in principle derive from parameters

  Would need to know the "One True Model"
- We derived a consistent non-parametric estimator, based on adaptive histograms[2]

  May not be an optimal estimator — but it's the first
- Using an estimated β_a complicates formulas but won't change the basics

[2] McDonald, Shalizi, and Schervish (2011), *Estimating beta-mixing coefficients via histograms*

Measuring capacity

How do we measure model capacity?

There are lots of ways of doing this!

Algorithmic Stability, Discrepancy, Covering/packing numbers, etc.

Most common in literature:

Rademacher complexity How well does the model seem to fit iid $\{+1, -1\}$ RVs?

$+$ Gives tightest bounds, don't have to use theory to calculate

$-$ Requires bounded loss functions

VC dimension Worst-case growth rate in covering number

All related, not quite the same

We use VC dimension

The why

- $+$ Fundamental: finite VC dimension is necessary and sufficient for learning with ergodic sources[3]
- $+$ Leads to distribution-free bounds (possibly more conservative than others)
- $+$ Works with unbounded loss functions
- $-$ Often very hard to find theoretically (heavy combinatorics)

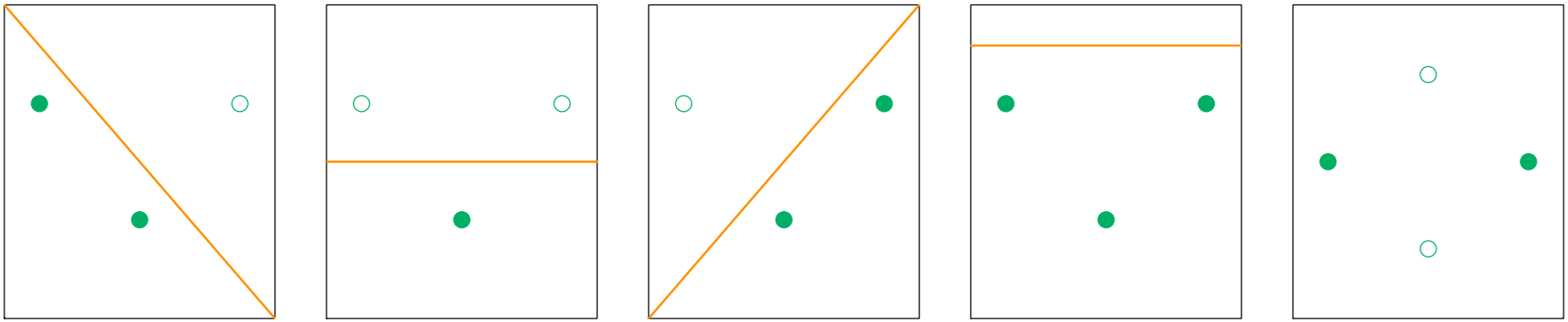
[3] Adams and Nobel (2010), *Uniform convergence of VC-classes under ergodic sampling*

The what: shattering

Definition (Vapnik and Chervonenkis (1971))

A collection of sets $\mathcal{D}$ shatters a finite set S when, for any $S' \subseteq S$, $S' = S \cap D$ for some $D \in \mathcal{D}$. [$\mathcal{D}$ can 'pick out' every subset S']

Let S be a set of points in $\mathbb{R}^2$. Let $\mathcal{D}$ be halfspaces in $\mathbb{R}^2$.
Then can shatter some 3-element sets, but no 4-element set.



The what: VC dimension

Definition (Vapnik and Chervonenkis (1971))

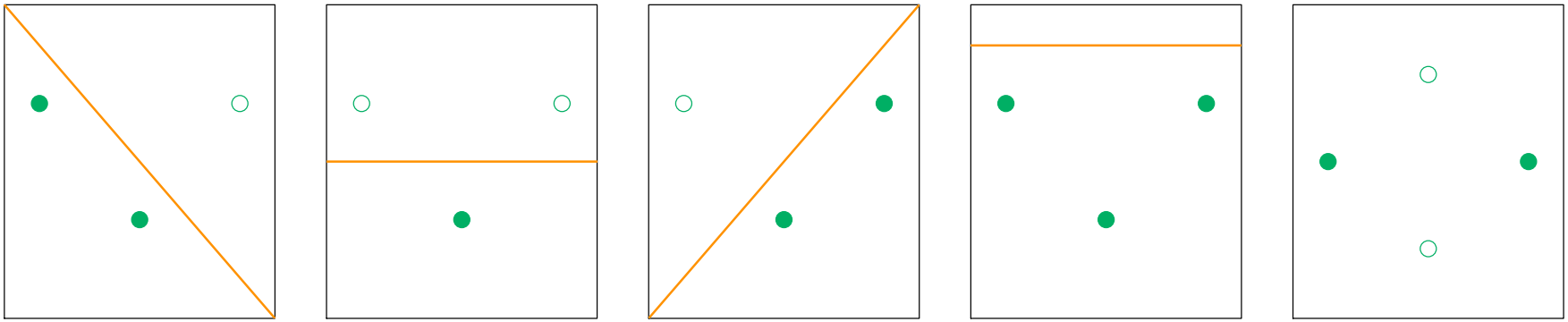
- The VC dimension of $\mathcal{D}$ is the size of the largest set it shatters.
- The VC dimension of a class of indicator functions is the VC dimension of the corresponding sets.
- The VC dimension of a class of real-valued functions is that of their collection of level sets.

Growth function of a collection of sets/functions = number distinguishable with n observations — Grows like 2^n

$$G(n, \mathcal{D}) \leq (n+1)^{\text{VCD}(\mathcal{D})}$$

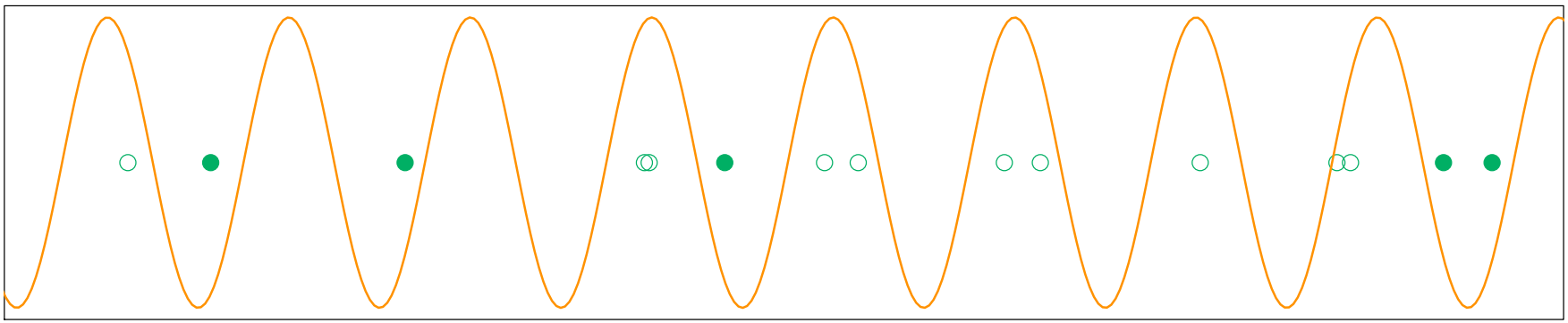
hence "dimension"

For p-dimensional linear models (with intercept), $\text{VCD} = p + 1$



In general $\text{VCD} \neq$ number of degrees of freedom

$\mathcal{D} = \{x \mapsto \sin(\omega x) : \omega \in \mathbb{R}\}$ has 1 parameter but $\text{VCD}(\mathcal{D}) = \infty$



Central result

Risk for time series learning

- Different from the IID case: observe data $Y_1^n := (Y_1, \ldots, Y_n)$.
- Fixed- vs. growing- memory predictors: can we ignore everything before the most recent d observations (AR) or not (MA, ARMA, state-space)?
- Leads to a slightly different definition of empirical risk

$$\widehat{R}_n(f) = \frac{1}{n-1} \sum_{i=1}^{n-1} \ell(f_i(Y_1^i), Y_{i+1})$$

f_i is defined to produce the prediction with data up to time i

- Generalization risk is the same

$$R_n(f) = \mathbb{E}\left[\ell(f(Y_1^n), Y_{n+1})\right]$$

Moment assumption

- Additive bounds rely on bounded losses: $\forall f \in \mathcal{F}$, and $\forall (x, y)$, $\ell(f(x), y) < M$ or the existence of an envelope.
- Can also bound a moment.
- Key assumption:[5] $\forall f \in \mathcal{F}$,

$$Q_n(f) := \sqrt{\mathbb{E}_{\mathbb{P}}\left[\ell(f(Y_1^n), Y_{n+1})^2\right]} \leq M < \infty$$

[[Strictly weaker than usual distributional assumptions on noise]]

[5] Cortes, Mansour, and Mohri (2010), "Learning bounds for importance weighting", *NIPS*

IID RESULT

Under this assumption,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_n(f)}{Q_n(f)}>\epsilon\right)\leq 4(2n+1)^{\mathrm{VCD}(\mathcal{F})}e^{-nr(\epsilon)/4}$$

Here $r(\epsilon)=\exp\left(W\left(-\frac{2\epsilon^2}{e^4}\right)+4\right)<\frac{\epsilon^{8/3}}{4^{2/3}}$

Putting the pieces together

1. Use IID results to bound deviation for each f
2. Use mixing to find out how much information is in the data
3. Use VC dimension to measure the capacity of the model
4. Result: bounds on generalization error (possibly including correction for growing memory)

Main theorem and its interpretation

Theorem (McDonald et al., 2011)

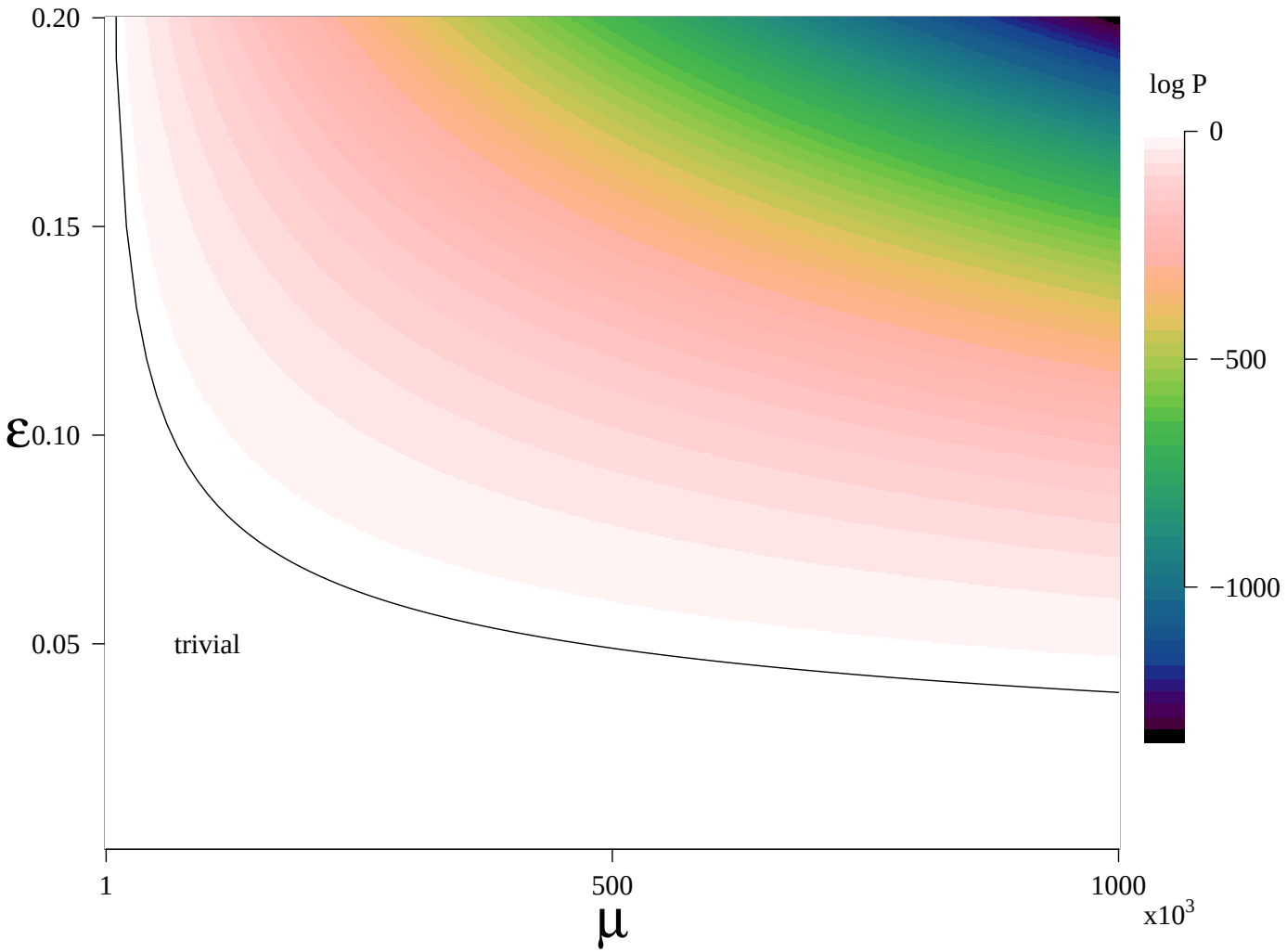
Assume mixing, the moment bound, and that $\mathcal{F}$ has fixed memory length d. Choose integers μ, a s.t. $2\mu a + d \leq n$ and $0 < \epsilon \leq 1$. Then

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{Q_n(f)} > \epsilon\right)$$
$$\leq 8(2\mu + 1)^{\mathrm{VCD}(\mathcal{F})} e^{-nr(\epsilon)/4} + 2\mu\beta_{a-d}$$

Meaning: with high probability, all the predictors in $\mathcal{F}$ come ϵ-close to their true performance after this much data

$\therefore$ with high probability $\widehat{f}$ will do no worse than this

Probability of maximum relative error exceeding ϵ



Inverting

- Invert by demanding confidence and finding precision:
- if $\eta > 2\mu\beta_{a-d}$,
- then with probability at least $1 - \eta$,
- simultaneously for all f (including $\widehat{f}$),

$$R_n(f) \leq \widehat{R}_n(f) + M\sqrt{\frac{\mathcal{E}_{\mathcal{F}}(4 - \log \mathcal{E}_{\mathcal{F}})}{2}}$$

with

$$\mathcal{E}_{\mathcal{F}} = \frac{4\,\mathrm{VCD}(\mathcal{F})\log(2\mu + 1) + \log 8/\eta'}{\mu}$$

$$\eta' = \eta - 2\mu\beta_{a-d}$$

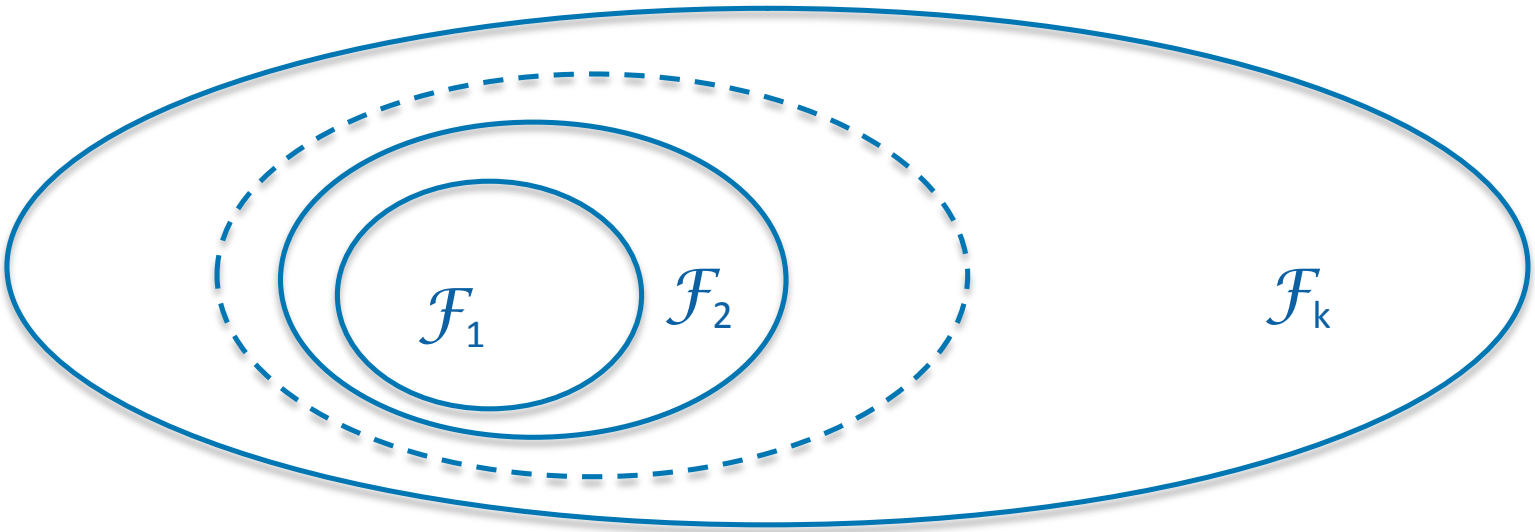
Extensions

Model selection

Multiple models $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k, \ldots$, with different capacities, and minimizers $\widehat{f}_k$ (assume some conditions)
Typical model selection (AIC, BIC, etc.):

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \widehat{R}_n(\widehat{f}_k) + p_k \lambda(n)$$

These work asymptotically at best



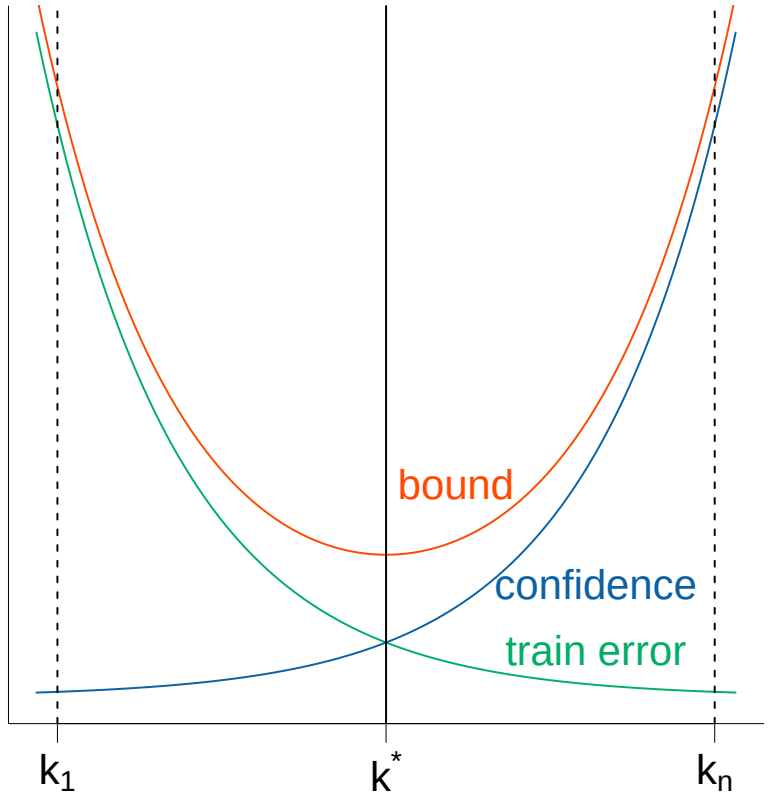
Model selection

Instead use structural risk minimization:

$$\widehat{k} = \operatorname*{argmin}_k \widehat{R}_n(\widehat{f}_k) + \\ + M\sqrt{\frac{\mathcal{E}_{\mathcal{F}_k}(4 - \log \mathcal{E}_{\mathcal{F}_k})}{2}}$$

Has nice properties:

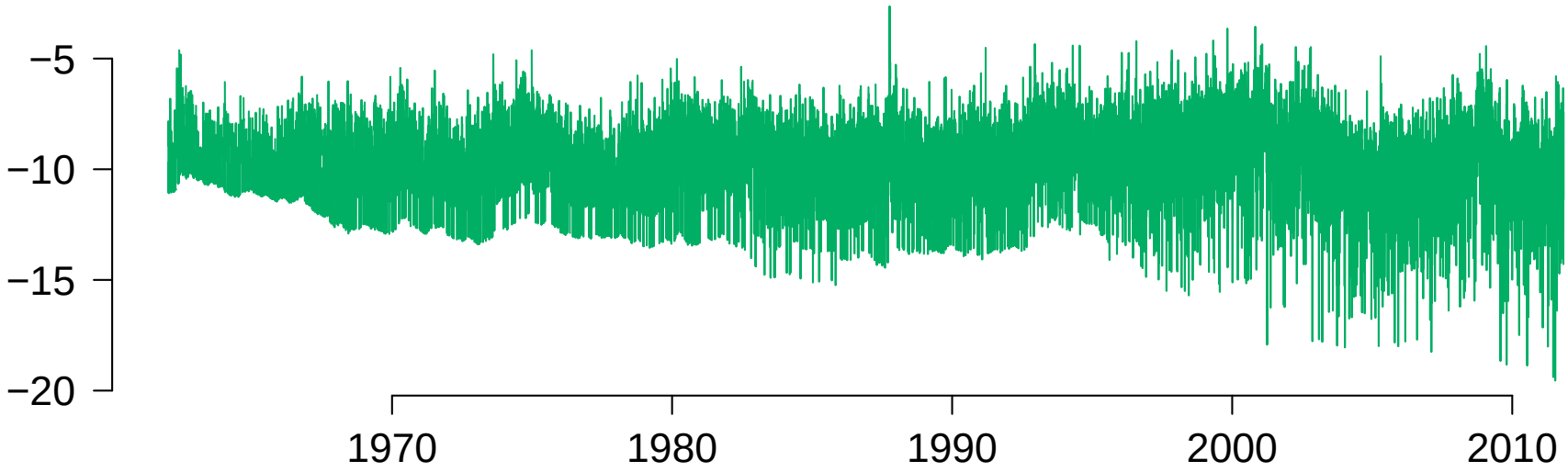
- picks out the best-predicting model with high probability
- risk-consistent in infinite-dimensional problems,
- etc.



Source: Vapnik (1998), *Statistical Learning Theory*, or Massart (2007) *Concentration inequalities and model selection*

A Small Worked Example

Daily log volatility for IBM, January 1962–October 2011



$n = 12541$, but $\mu = 846$, $a = 7$ due to dependence

Model	Training error	AIC-Baseline	Risk bound ($1 - \eta > 0.85$)	VCD
SV	1.82	-1124	4.89	3*
AR(2)	1.88	-348	3.26	3
Mean	1.91	0	2.81	1

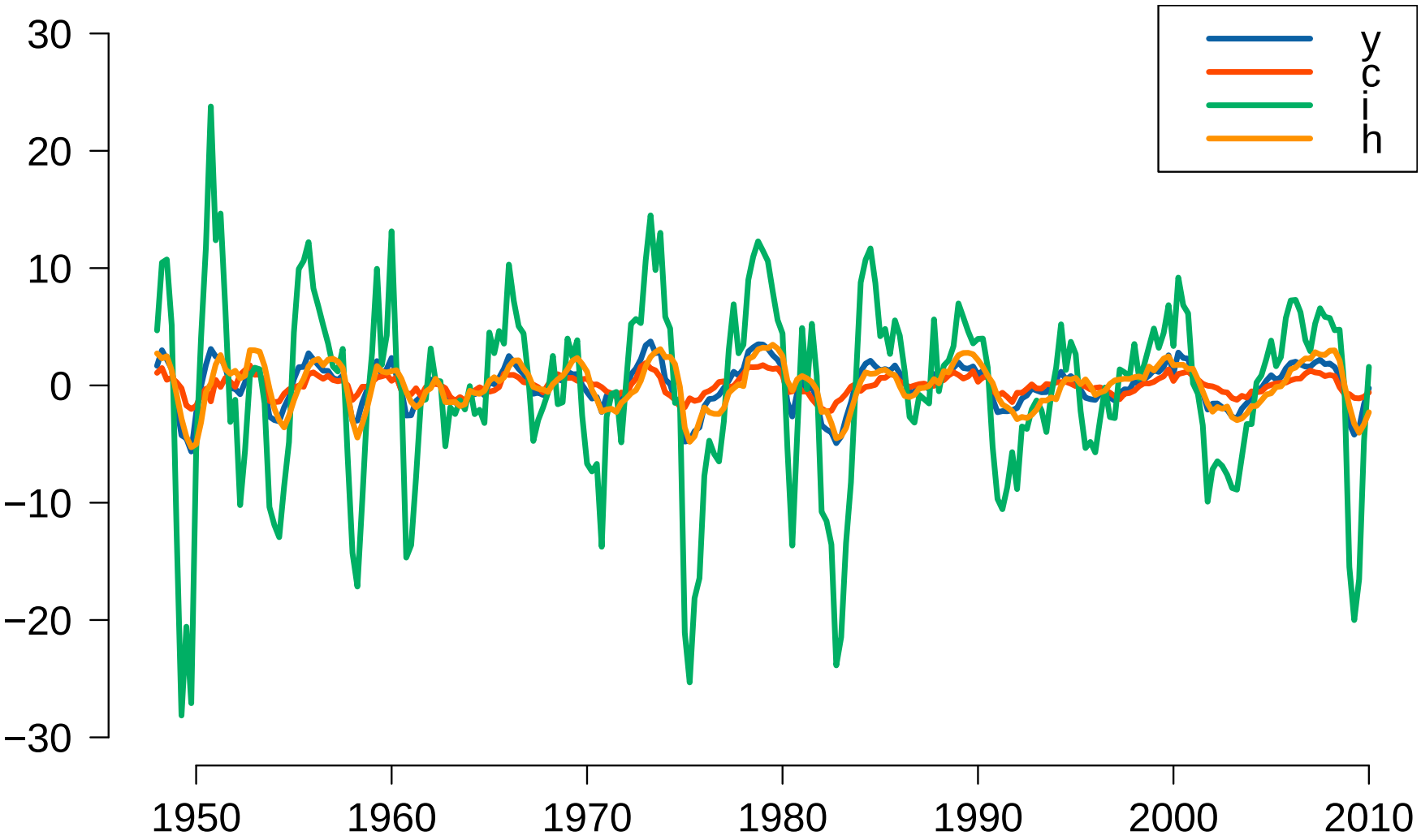
What about a DSGE?

$$\max_{\mathbf{c},\mathbf{l}} U = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left(\frac{c_t^{\varphi} l_t^{1-\varphi}}{1-\phi} \right)^{1-\phi}$$

subject to

$$\begin{aligned}
y_t &= z_t k_t^{\alpha} h_t^{1-\alpha}, \\
1 &= h_t + l_t, \\
y_t &= c_t + i_t, \\
k_{t+1} &= i_t + (1-\delta) k_t, \\
\ln z_t &= (1-\rho) \ln \bar{z} + \rho \ln z_{t-1} + \epsilon_t, \\
\epsilon_t &\overset{iid}{\sim} \mathrm{N}(0, \sigma_{\epsilon}^2).
\end{aligned}$$

What about a DSGE?



What about a DSGE?

$n = 249$

Estimated mixing coefficients imply $\mu = 24$

Bound is HUGE

The sample size is too small to provide confidence in complicated models

Mixing coefficients are not the culprit. Here I am being generous. Looking at correlation time of the (detrended) time series for GDP, there are about 12 "independent" data points

How tight are the bounds?

Assume $\beta_a = O(\exp(-a^\kappa))$.

Assume some other stuff.

Then, for suitably large n,

$$c\sqrt{\frac{\text{VCD}}{n}} \leq R(\widehat{f}) - R(f^*) \leq C\sqrt{\frac{\text{VCD}}{n^{\kappa/(1+\kappa)}}}$$

Constants are murder with small sample sizes.

Recapitulation

1. Assume stationary mixing data and a moment bound
2. Then we can use mixing to say how much information we have
3. And measure VC dimension to find the capacity of the model
4. And bound how optimistic the training error is as an estimate of the risk
5. The bounds hold for finite n
   and for mis-specified models
   and for all data sources

Summary

- Bounding generalization error is a sound and objective way to evaluate mis-specified predictive models
- We established how to do it for time-series data and time-series models
- Bounds shrink as you get more data and grow as models become more flexible
- All you have to do is run the calculations
- There are lots of ways to extend this, and even more to apply it

The end

Further Directions

- More direct treatment of infinite-memory case
- Other notions of weak dependence, beyond β-mixing
- Other notions of model capacity, beyond VC dimension, especially Rademacher complexity[6]
- Sharper, data-dependent bounds (e.g., coverage guarantees for stationary bootstraps?)
- Panel data
- Bounding regret rather than risk

[6] McDonald, Shalizi, and Schervish (2011), *Risk bounds without strong mixing*

Estimating β_a:

$$\beta_a = \int |p(x,y) - p_{-\infty:0}(x)p_{a:\infty}(y)| \, dxdy$$

Approximate via finite-length blocks

$$\beta_a^{(d)} = \int \left| p^{(d)}(x,y) - p_{-(d-1):0}(x)p_{a:(a+d)}(y) \right| dxdy$$

Using adaptive histograms, can consistently estimate both densities and do integral trivially

Let d grow at a rate just below $o(\log n)$ to get consistency,

$$\widehat{\beta_a^{(d)}} \to \beta_a$$

assuming only $\beta_a \to 0$ as $a \to \infty$

More on tightness of bounds

- Bounds are loose because they hold for potentially unlikely, truly awful distributions
- Bootstrap technique may give something tighter, more data dependent

- To get the upper/lower bound on Slide 54
  1. Assume bounded loss
  2. Exists N, st, $n > N$, $\exists c, C$
  3. c and C are independent of VCD, n
- If assume $\beta_a = O(a^{-r})$, then rate is same with $0 < \kappa < \frac{r-1}{2}$
- In DSGE, with estimated mixing coefficients let $\kappa \to \infty$

Stochastic volatility model

The SV model is typically given as

$$y_t = \tau z_t \exp(\rho_t/2), \qquad z_t \sim \mathrm{N}(0,1),$$
$$\rho_{t+1} = \phi\rho_t + w_t, \qquad w_t \sim \mathrm{N}(0,\sigma_\rho^2),$$

To estimate,

1. Transform to (linear) state space form by squaring and taking logs

   of the first (observation) equation
2. Predict $\log y_t^2$
3. Approximate the "growing memory model" with a fixed memory model $d = 2$

   hence VC dimension is no larger than 3
4. Include fudge factor to calculate the bounds

Relationship to state space models

DSGE Model

$$\max_{c_t, l_t} U = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t, l_t)$$

$$y_t = z_t q(k_t, n_t)$$

$$1 = n_t + l_t$$

$$y_t = c_t + i_t$$

$$k_{t+1} = i_t + (1 - \delta) k_t$$

$$z_t \sim \text{AR}(1)$$

Math games

State Space Model

$$\mathbf{x}_t = g(\boldsymbol{\alpha}_t, \epsilon_t)$$

$$\boldsymbol{\alpha}_{t+1} = h(\boldsymbol{\alpha}_t, \eta_{t+1})$$

$$\boldsymbol{\alpha}_1 \sim F$$

What About Specification Searches?

You published $\mathcal{F}$
but your theory didn't really pick it out
so you also tried $\mathcal{G}$ and $\mathcal{H}$
Our bound will then be overly optimistic
But an honest bound would just use the capacity of $\mathcal{F} \cup \mathcal{G} \cup \mathcal{H}$
Can be pushed further by using more information about the search process

Rademacher complexity

Definition

Define the Rademacher complexity of a function class $\mathcal{F}$ as

$$\mathfrak{R}(\mathcal{F}) = \mathbb{E}_X \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right| \right],$$

where σ_i are iid and $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$.

- Measures the maximum covariance between the predictions and random noise—how closely can some $f \in \mathcal{F}$ fit garbage?
- Removing $\mathbb{E}_X$ gives empirical Rademacher complexity
- $+$ Gives parametric rates if bounded loss, regularized objective
- $-$ Is ∞ if not bounded loss

Bibliography

- Adams, T., and Nobel, A. (2010), "Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling," The Annals of Probability, **38**(4), 1345–1367.
- Massart, P. (2007), "Concentration inequalities and model selection," in Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003, Springer.
- McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011a), "Estimated VC dimension for risk bounds," submitted for publication, arXiv:1111.3404.
- McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011b), "Estimating β-mixing coefficients," in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, eds. G. Gordon, D. Dunson, and M. Dudík, vol. 15, JMLR W&CP, arXiv:1103.0941.
- McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011c), "Estimating β-mixing coefficients via histograms," submitted for publication, arXiv:1109.5998.
- McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011d), "Risk bounds for time series without strong mixing," submitted for publication, arXiv:1106.0730.
- Smets, F., and Wouters, R. (2007), "Shocks and frictions in US business cycles: A Bayesian DSGE approach," American Economic Review, **97**(3), 586–606.
- Vapnik, V. (1998), Statistical learning theory, John Wiley & Sons, Inc., New York.
- Yu, B. (1994), "Rates of convergence for empirical processes of stationary mixing sequences," The Annals of Probability, **22**(1), 94–116.