Nonparametric risk bounds for time series prediction

Thesis Defense

Daniel McDonald

Committee:

Cosma Shalizi, Mark Schervish, Alessandro Rinaldo,
Larry Wasserman, and David N. DeJong

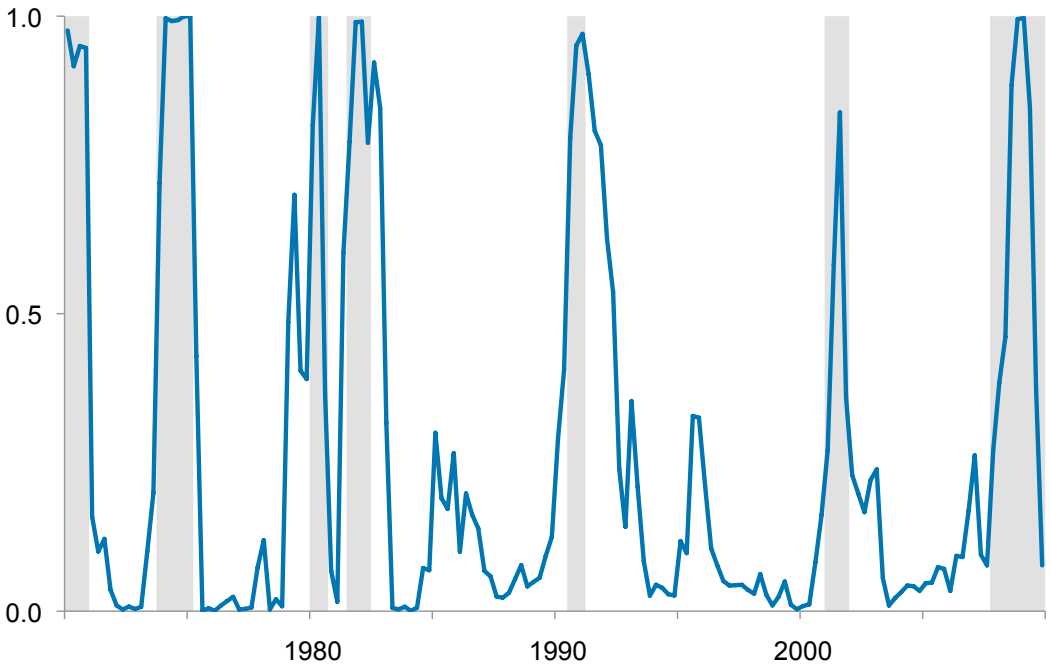
April 6, 2012

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

ECONOMIC FORECASTING

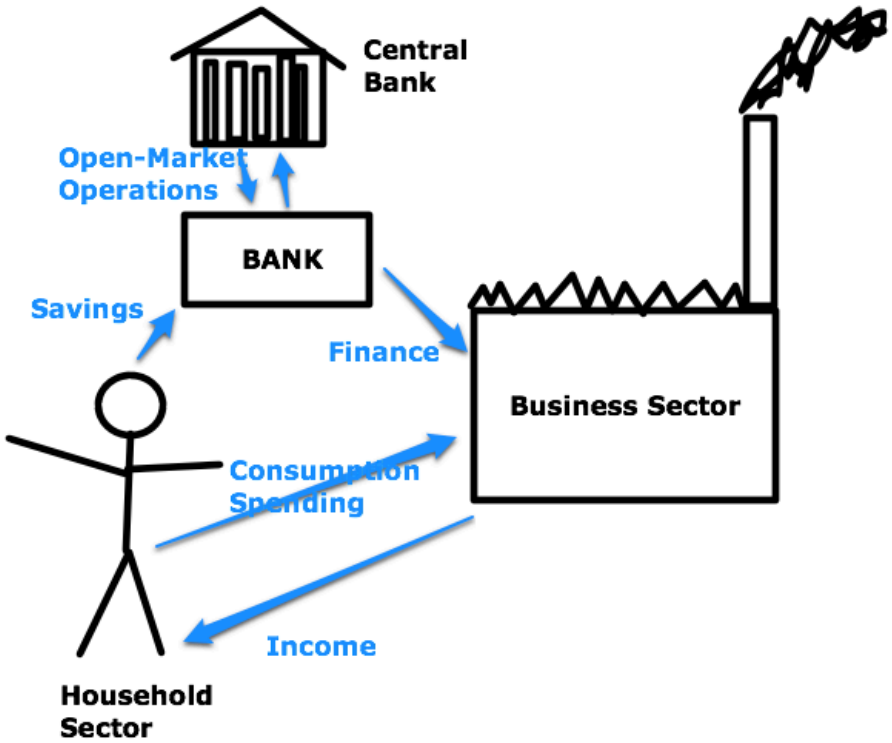


- ARIMA, ARFIMA, GARCH, etc.
- Dynamic Factor Models (Hamilton, Chib, Kim and Nelson, others)
- Systems of Equations models
- Dynamic Stochastic General Equilibrium (DSGE) models

Source: Econbrowser Recession Probabilities

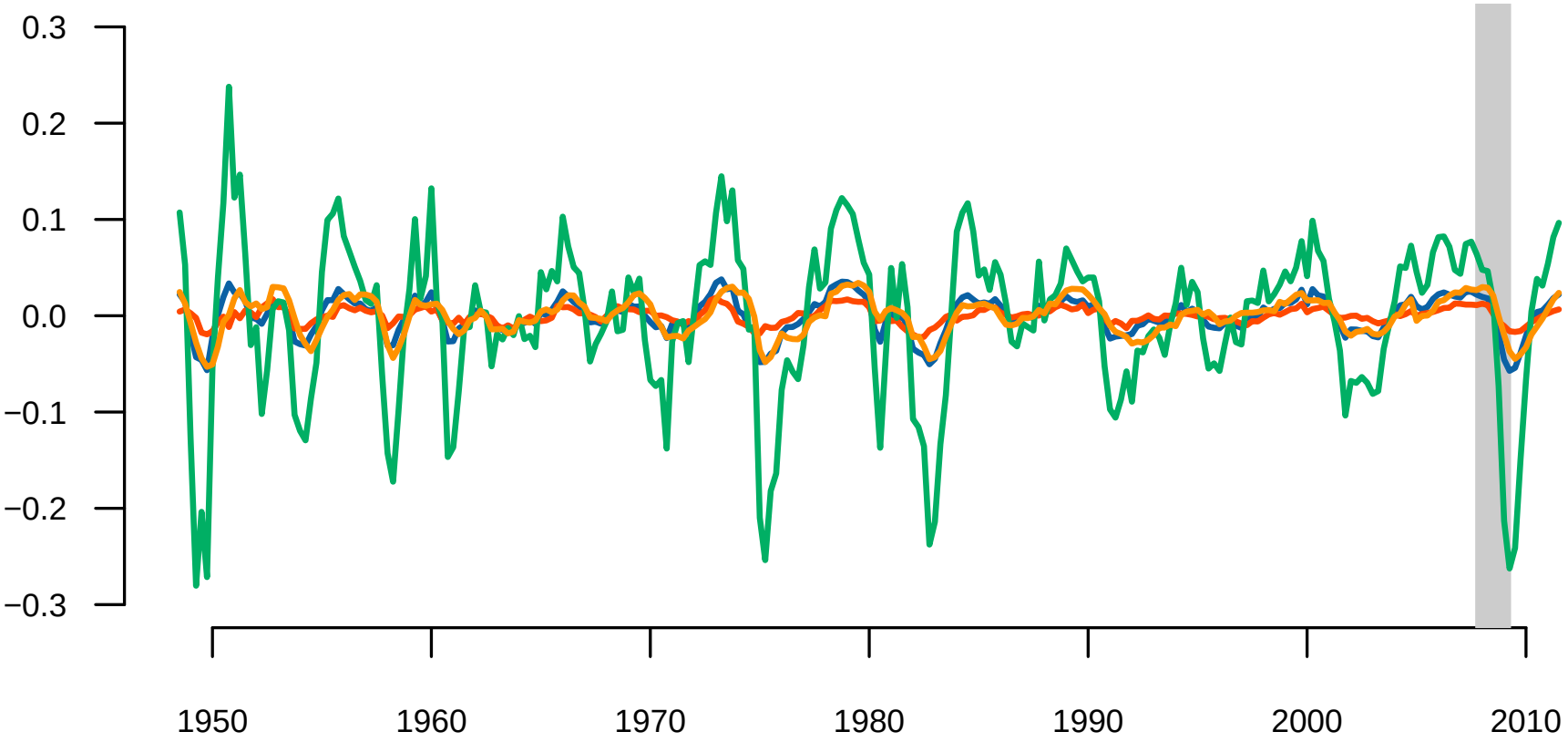
DSGE MODELS

- Most active area of macroeconomic research in the last 30 years
- Arose in response to the Lucas (1976) critique
- Pioneered by Kydland and Prescott (1982)
- Attempt to incorporate "rational behavior" into forecasting models



Source: Brad DeLong's realization of Daniel Davies' DSGE model

Some data (1948:I–2011:IV)



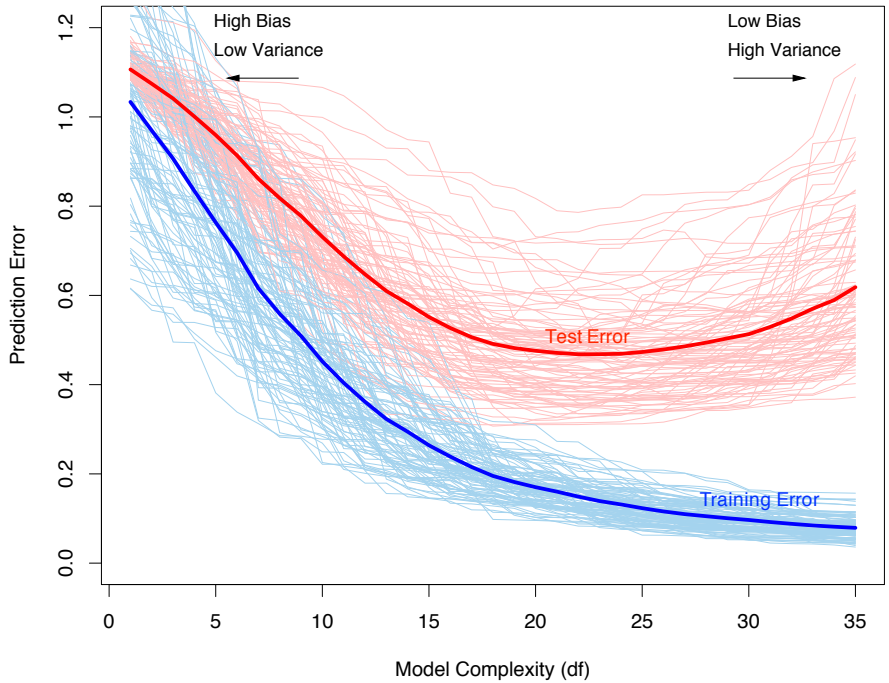
Income Consumption Investment Hours worked

MSE: In-sample .64 During recession 1.34

How to improve models??

- DSGEs came under fire for being unable to forecast the financial collapse of 2007–?

  Other models didn't either
- Solution in literature: add more 'stuff'
- Will lower in-sample error, but perhaps not out-of-sample error
- AIC, BIC, Bayes Factors, FPE, etc.



Source: Hastie, Tibshirani, and Friedman *The Elements of Statistical Learning*

THE BIG PICTURE

- Your favorite model often does worse out-of-sample than in-sample
- How much worse?
- Quantitative risk bounds provide insight
- The technology has mostly been for IID data and CS-style models
- We bound forecasting risk for time series and standard models
  - How much information do you really have?
  - How flexible is your model?
- This lets you assess your models rationally and objectively

Risk bounds: what and why?

Your favorite model fits the data pretty well

You'd like to know, with confidence, how well it will fit in the future

Risk

Risk of a function f for forecasting Y from X, with loss ℓ and data-source $\mathbb{P}$:

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

Why care about $R(f)$?

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

Comparison to other models

This is hard:

We don't know $\mathbb{P}$

If model was well-specified, could simulate

Models are rarely well-specified

Statistical learning theory to the rescue

Since the 1970s, and especially since the 1990s, statistics has figured out how to get confidence intervals for $R(f)$ which are

Distribution-free: hold uniformly over all $\mathbb{P}$
Agnostic: do not assume $\mathcal{F}$ is well-specified
Non-asymptotic: hold at finite n

This has helped move machine learning from a minor sub-field of AI to a major industrial technology
How does it work?
How can we use it with time series?

Statistical learning theory to the rescue

Since the 1970s, and especially since the 1990s, statistics has figured out how to get confidence intervals for $R(f)$ which are

Distribution-free: hold uniformly over all $\mathbb{P}$

Agnostic: do not assume $\mathcal{F}$ is well-specified

Non-asymptotic: hold at finite n

This has helped move machine learning from a minor sub-field of AI to a major industrial technology

How does it work?

How can we use it with time series?

Statistical learning theory to the rescue

Since the 1970s, and especially since the 1990s, statistics has figured out how to get confidence intervals for $R(f)$ which are

Distribution-free: hold uniformly over all $\mathbb{P}$
Agnostic: do not assume $\mathcal{F}$ is well-specified
Non-asymptotic: hold at finite n

This has helped move machine learning from a minor sub-field of AI to a major industrial technology
How does it work?
How can we use it with time series?

Statistical learning theory to the rescue

Since the 1970s, and especially since the 1990s, statistics has figured out how to get confidence intervals for $R(f)$ which are

Distribution-free: hold uniformly over all $\mathbb{P}$
Agnostic: do not assume $\mathcal{F}$ is well-specified
Non-asymptotic: hold at finite n

This has helped move machine learning from a minor sub-field of AI to a major industrial technology
How does it work?
How can we use it with time series?

Statistical learning theory to the rescue

Since the 1970s, and especially since the 1990s, statistics has figured out how to get confidence intervals for $R(f)$ which are

Distribution-free: hold uniformly over all $\mathbb{P}$
Agnostic: do not assume $\mathcal{F}$ is well-specified
Non-asymptotic: hold at finite n

This has helped move machine learning from a minor sub-field of AI to a major industrial technology
How does it work?
How can we use it with time series?

Summary of work contained herein

Estimating Mixing:

- McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011), "Estimating β-mixing coefficients."
- McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011), "Estimating β-mixing coefficients via histograms."

Bounds with different technology:

- McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011), "Generalization error bounds for stationary autoregressive models."
- McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011), "Risk bounds for time series without strong mixing."

Bounds for time series:

- McDonald, D. J., Shalizi, C. R., and Schervish, M. (2012), "Time series forecasting: model evaluation and selection using nonparametric risk bounds."

SUMMARY OF WORK CONTAINED HEREIN

Estimating Mixing:

MCDONALD, D. J., SHALIZI, C. R., AND SCHERVISH, M. (2011), "Estimating β-mixing coefficients."

MCDONALD, D. J., SHALIZI, C. R., AND SCHERVISH, M. (2011), "Estimating β-mixing coefficients via histograms."

Bounds with different technology:

MCDONALD, D. J., SHALIZI, C. R., AND SCHERVISH, M. (2011), "Generalization error bounds for stationary autoregressive models."

MCDONALD, D. J., SHALIZI, C. R., AND SCHERVISH, M. (2011), "Risk bounds for time series without strong mixing."

Bounds for time series:

- MCDONALD, D. J., SHALIZI, C. R., AND SCHERVISH, M. (2012), "Time series forecasting: model evaluation and selection using nonparametric risk bounds."

THE BASIC FORM OF STATISTICAL LEARNING THEORY

Get data $(x_1, y_1), \ldots, (x_n, y_n)$. Choose function class $\mathcal{F}$.

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:

$$\mathbb{P}\left(|\widehat{R}_n(f) - R(f)| > \epsilon \right) \leq \exp\left\{ -\frac{n\epsilon^2}{K^2} \right\}$$

Capacity

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

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

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,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}_n(f)| > \epsilon \right) \leq G(n, \mathcal{F}) \exp\left\{ -\frac{n\epsilon^2}{K^2} \right\}$$

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

Inverting

Invert to get confidence bounds

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

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

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

   What is $G(n, \mathcal{F})$?

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

Breeding dependent LLNs from independent ones

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

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?

- 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

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

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*

Moment assumption

- Additive bounds rely on bounded losses: $\forall f \in \mathcal{F}$, and $\forall (x, y)$, $\ell(f(x), y) < M$
- Unlimited losses have multiplicative bounds
- Key assumption:[5] for some $q > 2$, and $\forall f \in \mathcal{F}$,

$$\frac{\mathbb{E}_{\mathbb{P}}\left[\ell(f(Y_1^n), Y_{n+1})^q\right]^{1/q}}{R_n(f)} < M$$

Strictly weaker than usual distributional assumptions on noise

[5] Vapnik (1998), *Statistical learning theory*

VAPNIK'S IID RESULT

Under this assumption, then, with $\tau(q) = \sqrt[q]{\frac{1}{2}\left(\frac{q-1}{q-2}\right)^{q-1}}$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} > \epsilon\right) \leq 4GF(n,\mathcal{F}) \exp\left\{-\frac{n\epsilon^2}{4M^2\tau^2(q)}\right\}$$

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)

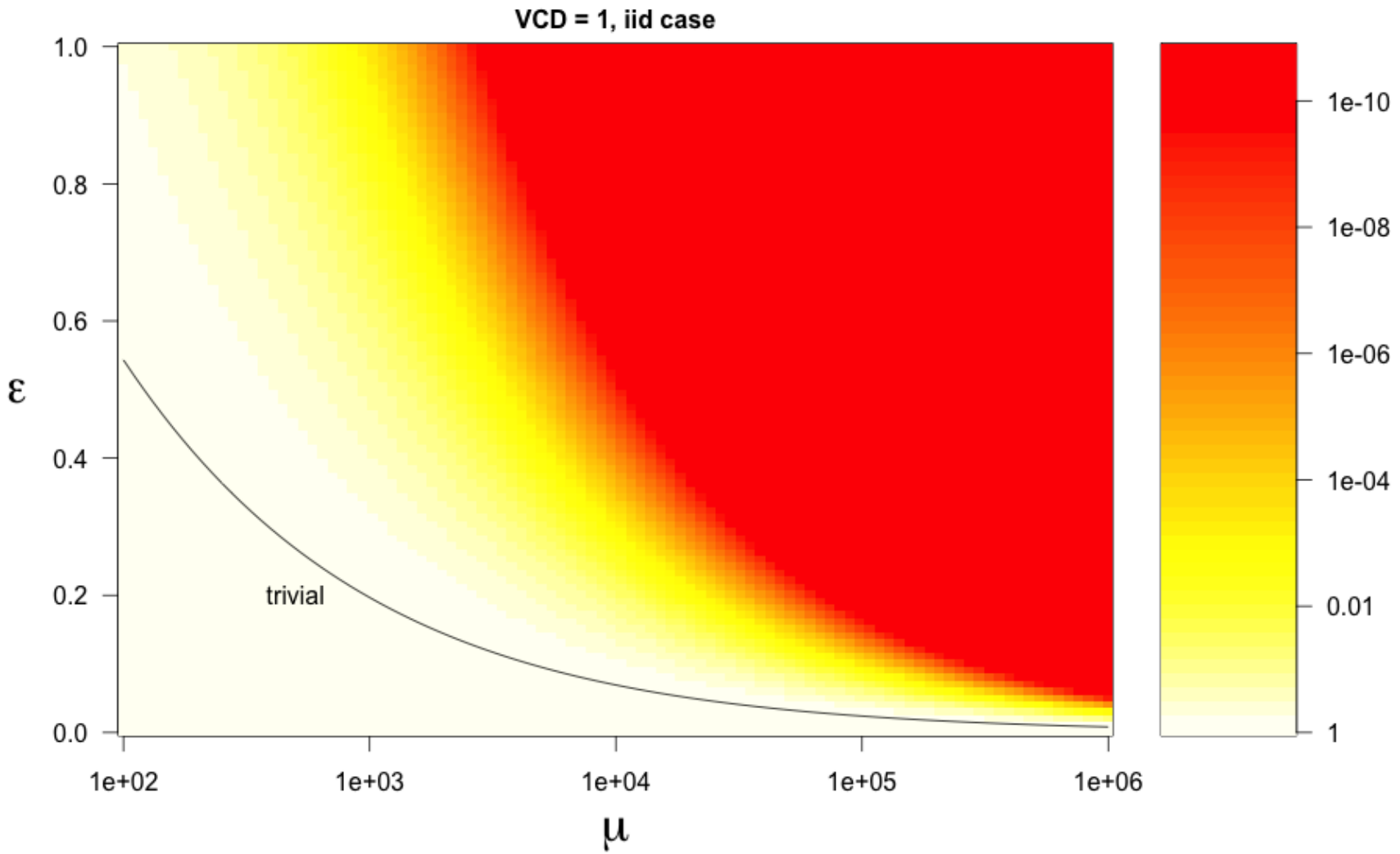
Assume mixing, the moment bound, and that $\mathcal{F}$ has fixed memory length d. Choose integers μ, a s.t. $2\mu a + d \leq n$ and $0 < \epsilon \leq 1$. Then

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} > \epsilon\right)$$
$$\leq 8GF(\mu, \mathcal{F}) \exp\left\{-\frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\} + 2\mu\beta_{a-d}$$

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 $\widehat{f}$),

$$R_n(f) \leq \widehat{R}_n(f) \times \frac{1}{(1 - \mathcal{E}(\mathcal{F}))_+}$$

with

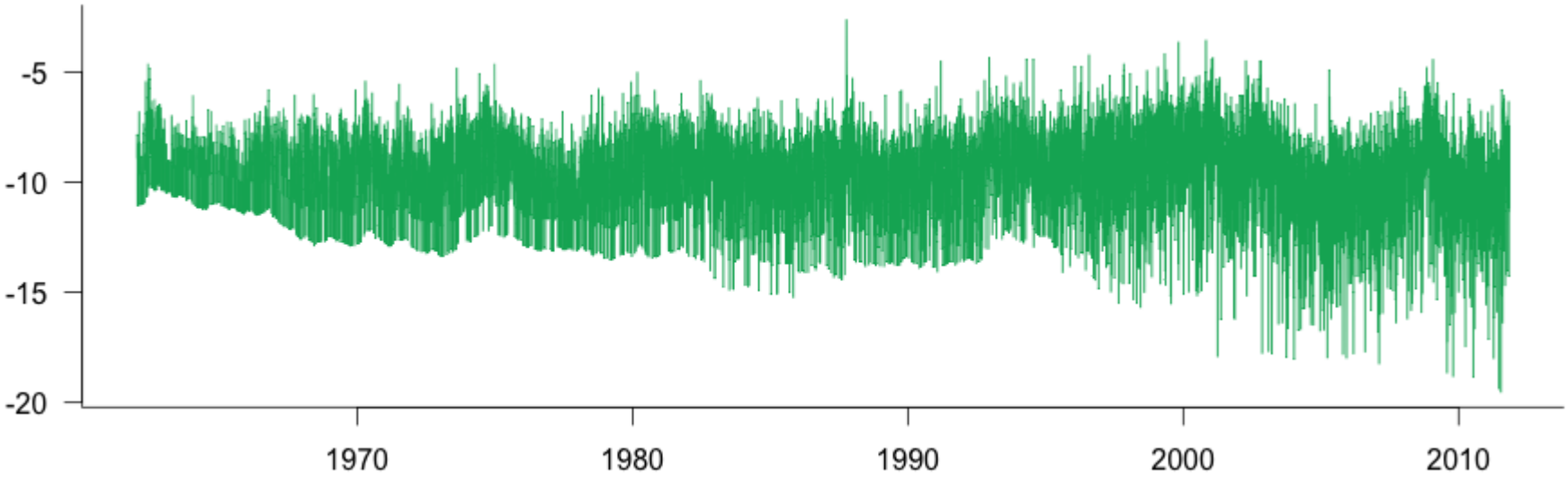
$$\mathcal{E}(\mathcal{F}) = 2M\tau(q)\sqrt{\frac{\log GF(\mu, \mathcal{F}) + \log 8/\eta'}{\mu}}$$

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

$$(u)_+ = \max(u, 0)$$

A Small Worked Example

Daily log volatility for IBM, January 1962–October 2011



$n = 12541$, but $\mu = 538\ (658)$, $a = 11\ (9)$ due to dependence

Model	Training error	AIC-Baseline	VCD	Risk bound ($1 - \eta > 0.85$)
SV	1.83	-2816	3*	16.68
AR(2)	1.88	-348	3	8.95
Mean	1.91	0	1	3.84

What about a DSGE?

$n = 255$

Estimated mixing coefficients imply $\mu = 31$

Bound is trivial $\Rightarrow \quad R_n(\widehat{f}) < \infty$

If data are IID, need $n > 481$ to get non-trivial bound

Using estimated mixing coefficients, need $n > 15000$!

Need data since before Moses left Egypt.

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

How tight are the bounds?

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

Assume some other stuff.

Then, for suitably large n,

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

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 use 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

Further Directions

- 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*

Summary

- Bounding generalization error is a sound and objective way to evaluate mis-specified predictive models
- I 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

Thanks for coming.

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 29
  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

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

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

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

YES! IT CONVERGES!

THE THEOREM

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\frac{R_n(f)-\widehat{R}_n(f)}{R_n(f)}>\epsilon\right)$$
$$\leq \quad 8GF(\mu,\mathcal{F})\exp\left\{-\frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\}+2(\mu-1)\beta_{a-d}$$

Suppose $\beta_a = o(a^{-r})$ for some $r > 0$. Can take $a_n = \Omega(n^{1/(1+r)})$
Then RHS = $o(n^{r/(1+r)})$.

Markov processes are known to have $\beta_a = o(\rho^{-a})$ for $\rho > 1$. Can take $a_n = o(n)$
Then RHS = $o(\min\{\rho, e\}^{-n})$.

Apart from some log terms

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.