Nonparametric risk bounds for time series prediction

Daniel McDonald

Department of Statistics
Carnegie Mellon University

http://www.stat.cmu.edu/∼danielmc

Joint work with:

Cosma Shalizi and Mark Schervish

January 12, 2012

Vignette

IF I COULD REARRANGE THE ALPHABET, I'D PUT YOUR SISTER AND I TOGETHER.
IS YOUR FATHER A THIEF? BECAUSE THAT'S TOTALLY MY JETTA YOU PARKED OUTSIDE.
YOU MUST BE TIRED, 'CAUSE YOU'VE BEEN RUNNING THROUGH MY MIND ALL NIGHT.
SCREAMING.

I have a model for predicting my success with pickup lines:

$$\mathbb{P}(\text{Yes}) = a_0 + a_1 \times \text{days since last shower} + a_2 \times \text{color of my shirt} + \cdots$$

I estimate my model with 29 data points

And find that it correctly predicts my success (or lack of) in 26 of 29 cases

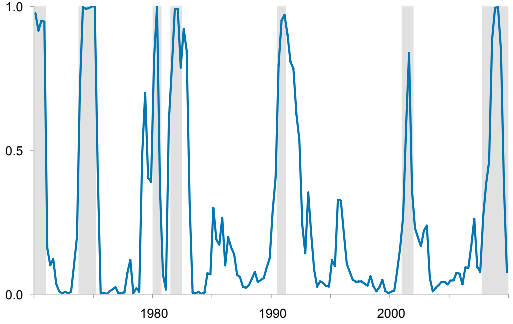
My model correctly predicts 3 of my next 8 attempts

Source: XKCD http://xkcd.com/279/

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 weird 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

Motivation: 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

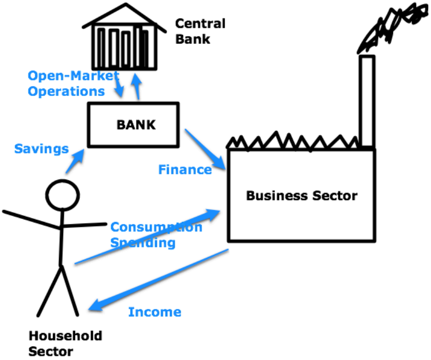
Source: Econbrowser Recession Probabilities

HOW TO IMPROVE MODELS??

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

  Other models didn't either
- Parameter explosion
- Smets and Wouters (2007) DSGE has ∼80 parameters and 7 latent variables

  Considered state of the art
- VARs: pk^2 parameters
- ∼250 data points



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

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 ν:

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

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 ν

If model was well-specified, could simulate

Models are rarely well-specified

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)

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 ν
- Agnostic: do not assume 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 ν

Agnostic: do not assume 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 ν
Agnostic: do not assume 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 ν
Agnostic: do not assume 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 ν
Agnostic: do not assume 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?

Plan of talk

1. Learning theory
2. Handling dependence
3. Measuring capacity
4. Our central result
5. Extensions and other considerations
6. Summary

The basic form of statistical learning theory

Get data $(x_1, y_1), \ldots, (x_n, y_n)$.
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:

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

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

Union bounds

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

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

Suppose $|\mathcal{F}|$ was finite.

And for each $f \in \mathcal{F}$,

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

Then, apply union bound to get

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

$|\mathcal{F}|$ 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,

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

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

Invert to get confidence bounds

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

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

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

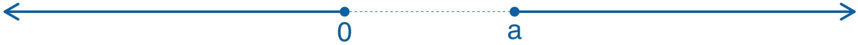
BREEDING DEPENDENT LLNS FROM INDEPENDENT ONES

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

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

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

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

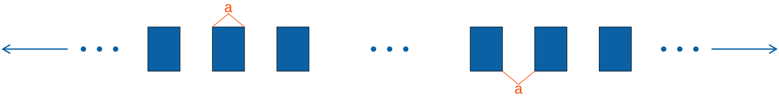


Intuition: at large separations, events are nearly independent

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, $\mathbb{P}_{\nu}(Z)$ with probabilities over IID blocks, $\mathbb{P}_{\widetilde{\nu}}(Z)$

   Then by a nice theorem,[1]

$$|\mathbb{P}_{\nu}(Z) - \mathbb{P}_{\widetilde{\nu}}(Z)| \leq \beta_a(\mu - 1)$$

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*

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)
- We show how to measure it accurately via simulation! [4]
- Can use the measurement instead of the theory in our main result with minor adjustments

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

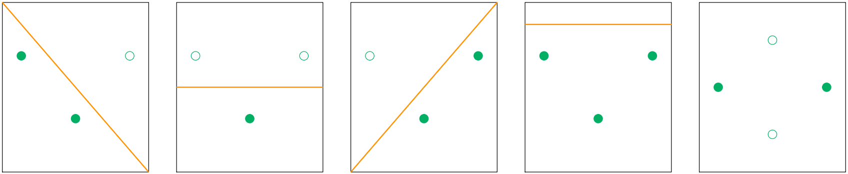
[4] McDonald, Shalizi and Schervish (2011), *Estimated VC-dimension for risk bounds*

The what: shattering

Definition (Vapnik and Chervonenkis (1971))

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

Let S be a set of points in $\mathbb{R}^2$. Let $\mathcal{C}$ 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))

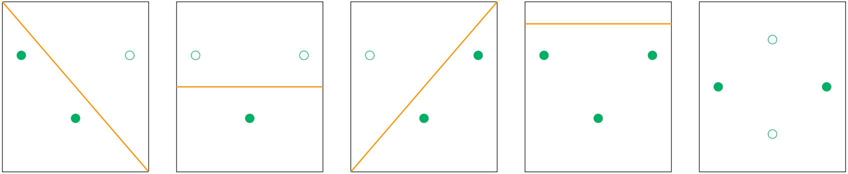
- The VC dimension of $\mathcal{C}$ is the size of the largest set it shatters.
- The VC dimension of a class of indicator functions is the VC dimension of the corresponding sets.
- The VC dimension of a class of real-valued functions is that of their collection of level sets.

Growth function of a collection of sets/functions = number distinguishable with n observations

$$G(n, \mathcal{C}) \leq \exp\left\{ \mathrm{VCD}(\mathcal{C}) \left(\log \frac{2n}{\mathrm{VCD}(\mathcal{C})} + 1 \right) \right\} = O(n^{\mathrm{VCD}(\mathcal{C})})$$

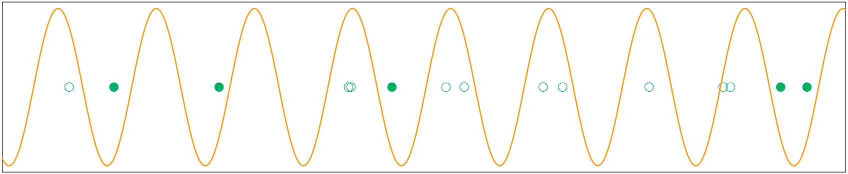
hence “dimension”

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



In general $\textsc{vcd} \neq$ number of degrees of freedom

$\mathcal{C} = \{x \mapsto \sin(\omega x) : \omega \in \mathbb{R}\}$ has 1 parameter but $\textsc{vcd}(\mathcal{C}) = \infty$



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 two slightly different notions of empirical risk

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

vs.

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

- 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 $\forall (x, y)$, $\ell(f(x), y) < M$
- Unlimited losses have to-within-a-factor bounds
- Key assumption:[5] for some $q > 2$, and $\forall f \in \mathcal{F}$,

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

Strictly weaker than usual distributional assumptions on noise

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

Vapnik's IID result

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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} > \epsilon\right) \\
&\leq \quad 4\exp\left\{\mathrm{VCD}(\mathcal{F})\left(\log\frac{2n}{\mathrm{VCD}(\mathcal{F})} + 1\right) - \frac{n\epsilon^2}{4M^2\tau^2(q)}\right\}
\end{aligned}$$

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)

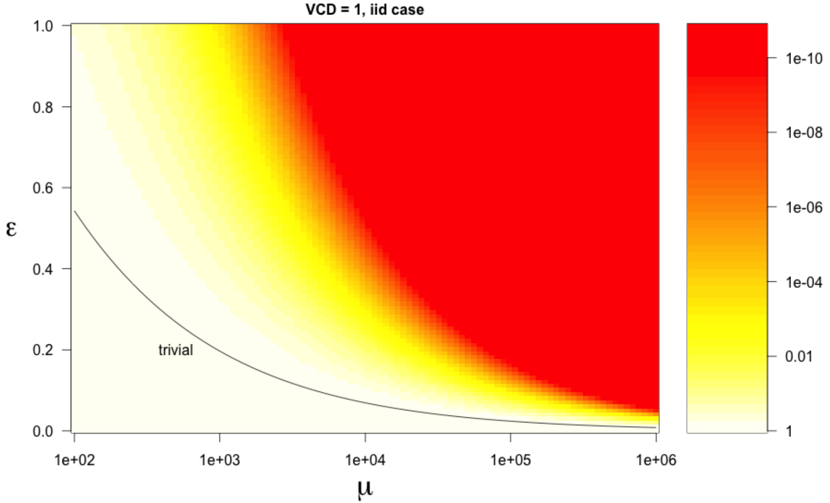
Assume mixing, the moment bound, that $\mathcal{F}$ has fixed memory length d, and that $\text{VCD}(\mathcal{F})$ is known. Choose integers μ, a s.t. $2\mu a + d \leq n$ and $0 < \epsilon \leq 1$. Then

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

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

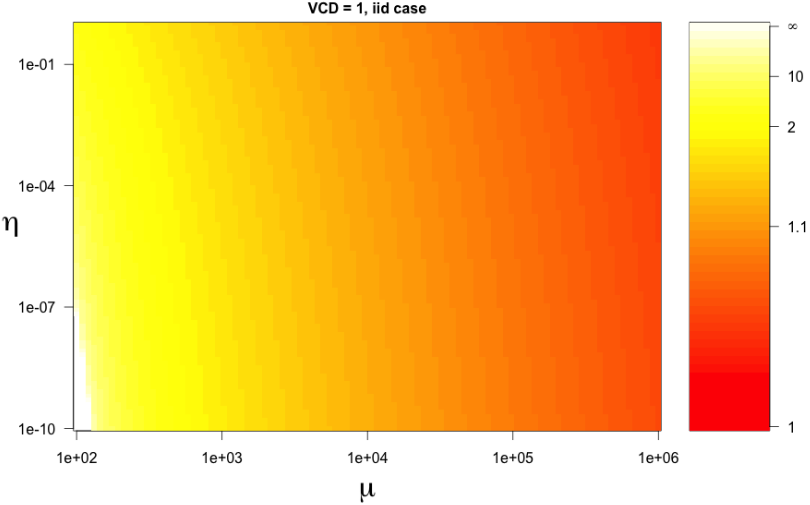
- Invert by demanding confidence and finding precision:
- if $\eta > 2(\mu - 1)\beta_{a-d}$,
- then with probability at least $1 - \eta$,
- simultaneously for all f (including $\widehat{f}$),

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

with

$$\mathcal{E}(\mathcal{F}) = \frac{2M\tau(q)}{\sqrt{\mu}}\sqrt{\text{VCD}(\mathcal{F})\left(\log\frac{2\mu}{\text{VCD}(\mathcal{F})} + 1\right) - \log(\eta'/8)}$$
$$\eta' = \eta - 2(\mu - 1)\beta_{a-d}$$
$$(u)_+ = \max(u, 0)$$

Size of the multiplicative penalty



Extensions

Don't know/can't find the VC dimension

Measure it: Add (shrinking) fudge factor to measured dimension and plug in, then add a little more probability of error.

Given arbitrary simulation time, the impact of the measurement goes away

Have a growing memory model

Linear case: Assume $\mathcal{F}$ is linear in the data.
Pick a finite-memory approximation order d, apply finite-memory theorem
add extra penalty to precision for the approximation

Penalty shrinks as d grows

WHAT ABOUT PRIORS AND REGULARIZATION?

- This is an entirely frequentist approach; real prior knowledge should be built into $\mathcal{F}$
- Could always use a prior as a regularization device, just like L^1 or L^2 penalties

  Again, models are mis-specified, so real degree of belief $= 0$
- Just measure capacity of the regularized model

  Adding bias to kill variance

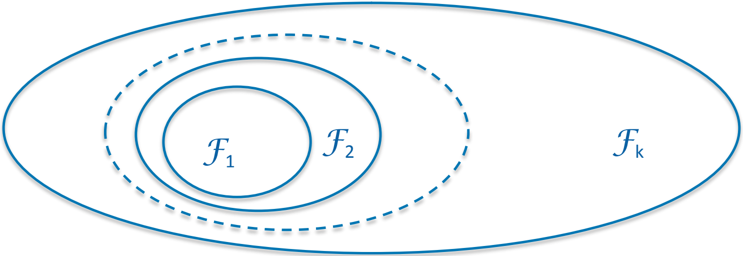
Model selection

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.):

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

These work asymptotically at best



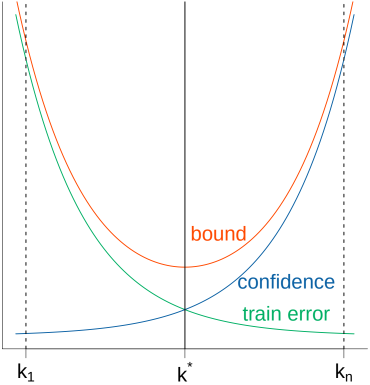
Model selection

Instead use structural risk minimization:

$$\widehat{k} = \operatorname*{argmin}_k \frac{\widehat{R}_n(\widehat{f}_k)}{(1 - \mathcal{E}(\mathcal{F}_k))_+}$$

Has nice properties:

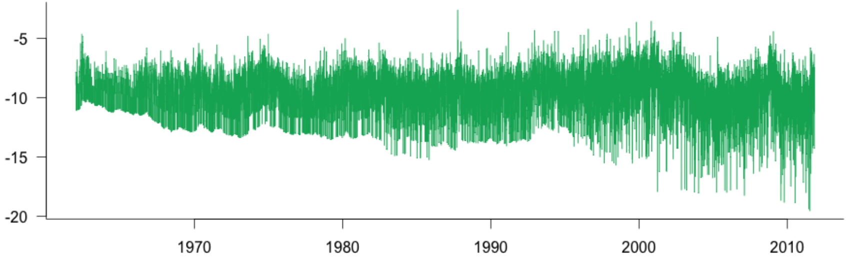
- picks out the best-predicting model with high probability
- risk-consistent in infinite-dimensional problems,
- etc.



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	9.81	3*
AR(2)	1.88	-348	5.37	3
Mean	1.91	0	3.46	1

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

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*

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

Thanks for inviting me.

ESTIMATING β_a:

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

Approximate via finite-length blocks

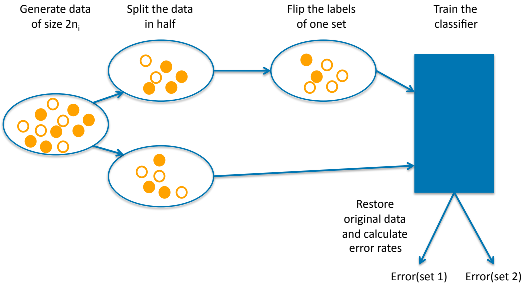
$$\beta_a^{(d)} = \int \left| p^{(d)}(x,y) - p_{-(d-1):0}(x) p_{a:(a+d)}(y) \right| dxdy$$

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 $a \to \infty$

ESTIMATING VC DIMENSION:

1. Pick a grid of design points (sample sizes) $n_1, \ldots n_k$, and repeat at each n_i:
2. 
Generate data of size $2n_i$
Split the data in half
Flip the labels of one set
Train the classifier
Restore original data and calculate error rates
Error(set 1)
Error(set 2)

   Calculate $|\text{Error(set 1) - Error(set 2)}|$.
   Average this discrepancy over m replications
3. Estimate VC dimension by nonlinear least-squares in a formula relating average discrepancy to n_i and VCD

WHEN THE BOUNDS ARE TOO LOOSE FOR COMFORT

US quarterly GDP goes back reliably to ≈ 1948

After de-trending, decay time implies ≈ 20 effectively-independent observations

Risk bound are <u>very wide</u> for any model

<u>This is not our fault</u>

This is the math's way of saying you do not have enough data

Without imposing very strong assumptions without support in <u>this</u> data

Might be supported by <u>other</u> data

the rational kernel hidden within the obscurantist shell of "calibration"

STOCHASTIC VOLATILITY MODEL

The SV model is typically given as

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

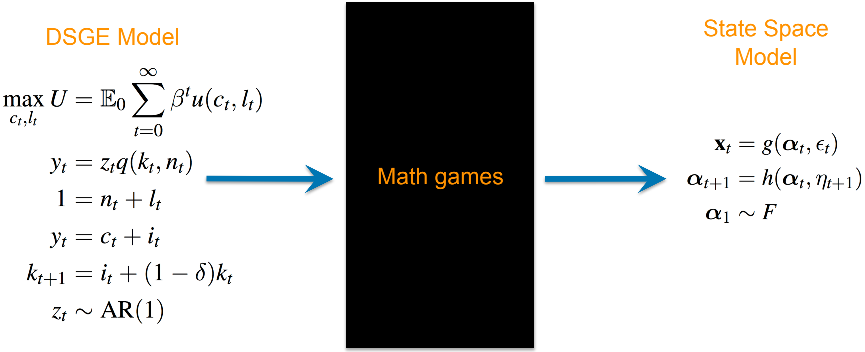
To estimate,

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

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

$$\begin{aligned} y_t &= z_t q(k_t, n_t) \\ 1 &= n_t + l_t \\ y_t &= c_t + i_t \\ k_{t+1} &= i_t + (1-\delta)k_t \\ z_t &\sim \mathrm{AR}(1) \end{aligned}$$

Math games

State Space Model

$$\begin{aligned} \mathbf{x}_t &= g(\boldsymbol{\alpha}_t, \epsilon_t) \\ \boldsymbol{\alpha}_{t+1} &= h(\boldsymbol{\alpha}_t, \eta_{t+1}) \\ \boldsymbol{\alpha}_1 &\sim F \end{aligned}$$

YES! IT CONVERGES!

THE THEOREM

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{R_n(f)} > \epsilon\right)$$

$$\leq \quad 8\exp\left\{\textsc{vcd}(\mathcal{F})\left(\log\frac{2\mu}{\textsc{vcd}(\mathcal{F})} + 1\right) - \frac{\mu\epsilon^2}{4M^2\tau^2(q)}\right\}$$

$$+2(\mu-1)\beta_{a-d}$$

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 [stat.ML].

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011c), "Estimating β-mixing coefficients via histograms," submitted for publication, arXiv:1109.5998 [math.ST].

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011d), "Risk bounds for time series without strong mixing," submitted for publication, arXiv:1106.0730 [stat.ML].

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.