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.

![Graph showing prediction error vs model complexity](image)

Source: Hastie, Tibshirani, and Friedman *The Elements of Statistical Learning*
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
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 $\ell$ and data-source $\mathbb{P}$:

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

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
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 $P$
- **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?
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?**
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?
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?
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:


Bounds with different technology:


Bounds for time series:

SUMMARY OF WORK CONTAINED HEREIN

Estimating Mixing:


Bounds with different technology:


Bounds for time series:

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):

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

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

Deviation inequalities for fixed functions:

\[
P\left(|\hat{R}_n(f) - R(f)| > \epsilon\right) \leq \exp \left\{-\frac{n\epsilon^2}{K^2}\right\}
\]
All well and good, but what about functions chosen using the data? Often select:

\[
\hat{f} := \arg\min_{f \in \mathcal{F}} \hat{R}_n(f) = \arg\min_{f \in \mathcal{F}} \left\{ R(f) + \gamma_n(f) \right\}
\]

**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,

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

Trade off precision [depends on \( \epsilon \)] and confidence [depends on \( n, \epsilon \)]
Invert to get confidence bounds

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

$$R(\hat{f}) \leq \hat{R}_n(\hat{f}) + K \sqrt{\frac{\log G(n, \mathcal{F}) + \log 1/\eta}{n}}$$
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 $\beta$-mixing (absolutely regular) process

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

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

$\beta$-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 \(2\mu\) blocks of length \(a\)
   Choose \(\mu, a\) 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, \(\tilde{\mathbb{P}}(Z)\)
   Then by a nice theorem, \(^1\)
   \[
   |\mathbb{P}(Z) - \tilde{\mathbb{P}}(Z)| \leq \beta_a \mu
   \]

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*
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 $\forall f \in \mathcal{F}$,

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

Strictly weaker than usual distributional assumptions on noise

$^5$ Vapnik (1998), Statistical learning theory
Vapnik’s IID Result

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

$$\mathbb{P}\left( \sup_{f \in \mathcal{F}} \frac{R_n(f) - \hat{R}_n(f)}{R_n(f)} > \epsilon \right) \leq 4GF(n, \mathcal{F}) \exp \left\{ -\frac{n\epsilon^2}{4M^2\tau^2(q)} \right\}$$
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 \( \mu, a \) s.t. \( 2\mu a + d \leq n \) and \( 0 < \epsilon \leq 1 \). Then

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

**Meaning:** with high probability, all the predictors in \( \mathcal{F} \) come \( \epsilon \)-close to their true performance after this much data.

\( \therefore \), with high probability \( \hat{f} \) will do no worse than this
PROBABILITY OF MAXIMUM RELATIVE ERROR EXCEEDING $\epsilon$
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 $\hat{f}$),

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

with

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

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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Training error</th>
<th>AIC-Baseline</th>
<th>VCD</th>
<th>Risk bound $(1 - \eta &gt; 0.85)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV</td>
<td>1.83</td>
<td>-2816</td>
<td>3*</td>
<td>16.68</td>
</tr>
<tr>
<td>AR(2)</td>
<td>1.88</td>
<td>-348</td>
<td>3</td>
<td>8.95</td>
</tr>
<tr>
<td>Mean</td>
<td>1.91</td>
<td>0</td>
<td>1</td>
<td>3.84</td>
</tr>
</tbody>
</table>
What about a DSGE?

\[ n = 255 \]

Estimated mixing coefficients imply \( \mu = 31 \)

Bound is trivial \( \Rightarrow R_n(\hat{f}) < \infty \)

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

Assume some other stuff.

Then, for suitably large $n$,

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

Constants are murder with small sample sizes.
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 $\beta$-mixing
- Other notions of model capacity, beyond VC dimension, especially Rademacher complexity
- 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*
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.
Thanks for coming.
**Estimating $\beta_a$:**

\[
\beta_a = \int |p(x, y) - p_{-\infty:0}(x)p_{a:\infty}(y)| \, dx \, dy
\]

Approximate via finite-length blocks

\[
\beta_{a}^{(d)} = \int |p^{(d)}(x, y) - p_{-(d-1):0}(x)p_{a:(a+d)}(y)| \, dx \, dy
\]

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,

\[
\hat{\beta}_{a}^{(d)} \rightarrow \beta_a
\]

assuming only $\beta_a \rightarrow 0$ as $a \rightarrow \infty$
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 $\text{vCD, } n$

- 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 $\kappa \to \infty$
The SV model is typically given as

\[ y_t = \tau z_t \exp(\rho_t/2), \quad z_t \sim N(0, 1), \]
\[ \rho_{t+1} = \phi \rho_t + w_t, \quad w_t \sim N(0, \sigma^2_\rho), \]

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

\begin{align*}
\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)
\end{align*}

State Space Model

\begin{align*}
x_t &= g(\alpha_t, \epsilon_t) \\
\alpha_{t+1} &= h(\alpha_t, \eta_{t+1}) \\
\alpha_1 &\sim F
\end{align*}
YES! IT CONVERGES!

THEOREM

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

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
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

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

where $\sigma_i$ 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 $\infty$ if not bounded loss


