Generalization Error Bounds for State Space Models

with an application to economic forecasting

Daniel McDonald

Department of Statistics
Carnegie Mellon University

http://www.stat.cmu.edu/∼danielmc

Joint Statistical Meetings

August 3, 2010

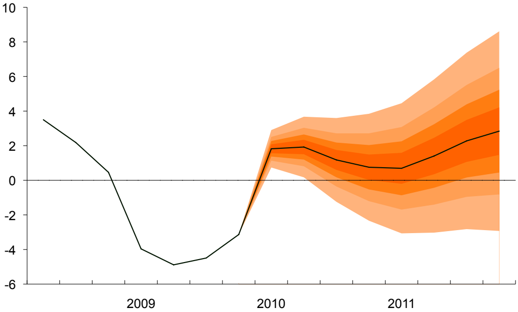
FORECASTING

- Given some data

$$x_1, \ldots, x_T \in \mathcal{X}$$

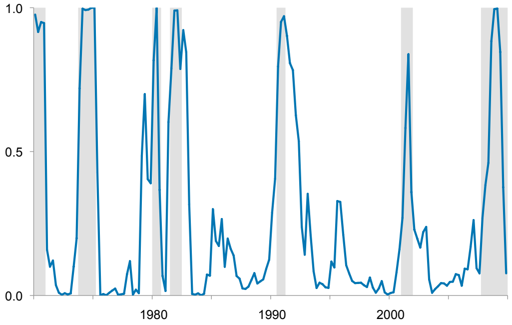
- Want to predict the next data point(s)

$$x_{T+1}, \ldots, x_{T+k}$$



Source: Czech National Bank

Methods of economic forecasting



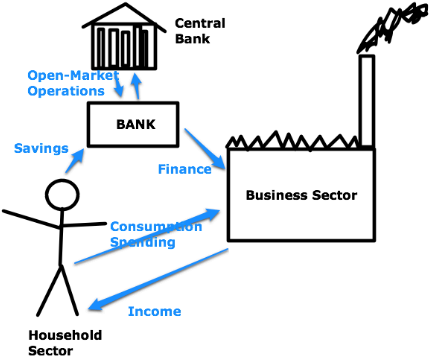
- VAR, ARIMA, GARCH
- Dynamic Factor Models (Hamilton, Chib, Kim and Nelson, others)
- Systems of Equations models
- Dynamic Stochastic General Equilibrium (DSGE) models

- All have equivalent representations as a state space model

Source: Econbrowser Recession Probabilities

Simplicity/Complexity

- Unclear if these models are "good"
- Lots of economic arguments Pro/Con
- What about statistical behavior?
- Overfit/Underfit
- How do predictions compare across different SS models?



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

Results from statistical learning

Robust comparisons/evaluations

Develop probabilistic bounds on the prediction error of state space models.

Forecasting framework

1. Observe training data $D_n = \{(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\}$ from some stochastic process μ
2. Choose model class $\mathcal{F}$ from which to construct predictors, e.g. AR(p), DSGE, regression, wavelets, Dynamic Factor models, etc.
3. Use a loss function $\ell(Y, f(X))$ to measure performance of candidate predictors $f \in \mathcal{F}$
4. Estimate the model using D_n, to produce $\widehat{f}$, your proposed forecasting model

GENERALIZATION ERROR

- Want to control the generalization error, or risk, of chosen predictor $\widehat{f}$

$$R(\widehat{f}) = \mathbb{E}_{\mu}[\ell(Y_0, \widehat{f}(X_0)) \mid D_n]$$

- But the stochastic process μ is unknown
- Usually estimate $R(\widehat{f})$ with training error

$$R_n(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i, \widehat{f}(X_i))$$

- Since $R(\widehat{f})$ is an expectation

$$R_n(\widehat{f}) = R(\widehat{f}) + \gamma_n(\widehat{f})$$

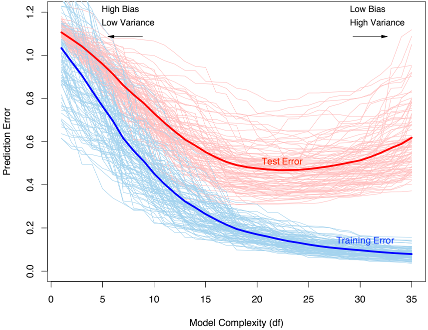
where $\gamma_n(\widehat{f})$ measures discrepancy between sample D_n and the true DGP

Training error

- Usually select

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} R_n(f)$$
$$= \operatorname*{argmin}_{f \in \mathcal{F}} [R(f) + \gamma_n(f)]$$

- Minimizing $R_n(f)$ conflates risk and in-sample noise
- So $\mathbb{E}_\mu[R_n(\widehat{f})] < R(\widehat{f})$
- Model comparisons using $R_n(\widehat{f})$ lead to choosing overly complex $\mathcal{F}$—overfitting



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

Error bounds

Risk

$$R(\widehat{f}) = \mathbb{E}_\mu[\ell(Y_0, \widehat{f}(X_0)) \mid D_n]$$

- Estimation of $R(\widehat{f})$ is a hard problem since μ is unknown
- Instead, derive probabilistic upper bounds
- These bounds depend on $\mathcal{F}$ — one needs to characterize the complexity of different function classes

Error bounds (cont.)

- Can derive upper bound

$$R(\widehat{f}) \leq R_n(\widehat{f}) + \max_{f \in \mathcal{F}} \gamma_n(f)$$

- We cannot calculate $\max_{f \in \mathcal{F}} \gamma_n(f)$, but we can bound it with high probability
- With probability at least $1 - \eta$,

$$\max_{f \in \mathcal{F}} \gamma_n(f) \leq \delta(C(\mathcal{F}), n, \eta).$$

- $C(\mathcal{F})$ characterizes the complexity of $\mathcal{F}$
- Many complexity measures — VC Dimension, covering numbers, algorithmic stability, and Rademacher complexity

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 correlation between the predictions and random noise — how closely can some $f \in \mathcal{F}$ fit garbage?
- Gives tight bounds
- Removing $\mathbb{E}_X$ gives empirical Rademacher complexity

- Bound the Rademacher complexity of the class of models

$$\mathcal{F}_p = \left\{ \varphi_1, \ldots, \varphi_p : X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t \text{ and } X_t \text{ is stationary} \right\}$$

- Stationarity requires the roots of $p(z) = z^p + \varphi_1 z^{p-1} + \cdots + \varphi_p$ lie inside the complex unit disc.
- Can show that a sufficient condition is[1]

$$||\varphi||_2^2 \leq \sum_{i=1}^{p} \binom{p}{i}^2 = \binom{2p}{p} - 1$$

[1] Fam and Meditch 1978

Bounds for stationary AR(p) models (cont.)

- This result + Bartlett and Mendelson 2002 + Mohri and Rostamizadeh 2009 = risk bound for loss functions $\ell < M$.

Bound for AR(p) models

With probability at least $1 - \eta$,

$$R(\widehat{f}) < R_a(\widehat{f}) + 2\sqrt{\frac{p}{n}}\sqrt{\left(\binom{2p}{p} - 1\right)\mathbb{V}X_1} + M\sqrt{\frac{\log 2/\eta'}{2a}}$$

- a and η' depend on the serial dependence
- a is like an effective sample size
- As $n \longrightarrow \infty$, $\eta' \longrightarrow \eta$ and $a \longrightarrow \infty$ if the serial dependence decays quickly enough
- Thus $R(\widehat{f}) - R_a(\widehat{f}) \xrightarrow{n\to\infty} 0$

Who cares?

- Bounds are good for policy makers
- Can communicate the likelihood of large forecasting mistakes
- Can use to robustly compare competing models, classes of models
- Can tell you how much data you need to fit that DSGE with 20 structural shocks and 100 parameters

THE END

Questions?

danielmc@stat.cmu.edu

Time series bounds

Theorem

Let $\mathcal{H}$ be the space of losses bounded above by M. Then given a sample from a stationary β-mixing distribution, for all $m, a > 0$ with $2ma = n$ and $\eta > 2(a-1)\beta(m)$, then for all $f \in \mathcal{F}$, with probability at least $1-\eta$,

$$R(f) < R_a(f) + \mathfrak{R}_a(\mathcal{H}) + M\sqrt{\frac{\log 2/\eta'}{2a}}$$

with $\eta' = \eta - 2(a-1)\beta(m)$.

Source: Mohri and Rostamizadeh 2009

Implications

Theorem

$$R(f) < R_a(f) + \mathfrak{R}_a(\mathcal{H}) + M\sqrt{\frac{\log 2/\eta'}{2a}}$$

- The effective sample size is not n but a
- The empirical risk is based on a data points separated by a distance $2m$
- Faster decay in $\beta(m)$ means more 'independent' samples, smaller third term
- Second term is Rademacher complexity of the loss space
- Can substitute empirical Rademacher complexity with slight modifications
- M is an upper bound for the loss

- Ordinary linear regressions can be written as kernel regressions. Let

$$\alpha_i = \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'\mathbf{Y}\right)_i$$
$$k(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{X}_i\mathbf{X}_j',$$

  where $\mathbf{X}$ is the $n \times p$ design matrix, $\mathbf{Y}$ are the responses, and $\mathbf{X}_i$ is the i^{th} row of the design matrix.
- Requiring $\sum_{i,j} \alpha_i\alpha_j k(\mathbf{X}_i, \mathbf{X}_j) \leq \gamma^2$
- Corresponds $||\widehat{\beta}^{OLS}||_2^2 \leq \gamma^2$, or ridge regression

$$\mathcal{F}_p \subseteq \overline{\mathcal{F}_p} = \left\{ \varphi_1, \ldots, \varphi_p : x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} \text{ and } ||\varphi||_2^2 \leq \binom{2p}{p} - 1 \right\}$$

Allows application of kernel regularized result[1]

$$\mathfrak{R}(\mathcal{F}_p) \leq \mathfrak{R}(\overline{\mathcal{F}_p}) \leq \frac{2}{\sqrt{n}} \sqrt{\left(\binom{2p}{p} - 1 \right) \mathbb{E}\mathbf{X_1}\mathbf{X_1}'}$$

$$\mathfrak{R}_n(\mathcal{F}_p) \leq \mathfrak{R}_n(\overline{\mathcal{F}_p}) \leq \frac{2}{\sqrt{n}} \sqrt{\left(\binom{2p}{p} - 1 \right) \frac{1}{n} \sum_{t=i}^{n} \mathbf{X_i}\mathbf{X_i}'}$$

[1] Bartlett and Mendelson 2002