Statistical Machine Learning: Economic Forecasting

Daniel J. McDonald

Indiana University, Bloomington

mypage.iu.edu/∼dajmcdon

February 24-26, 2015

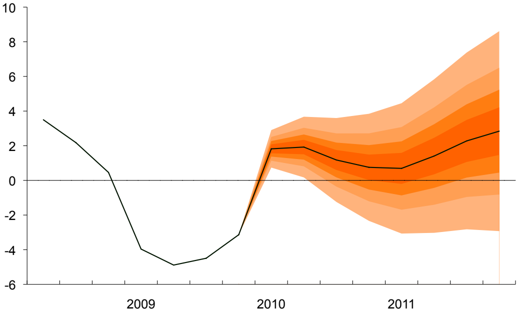
FORECASTING

- Given some data

$$y_1, \ldots, y_n \in \mathbb{R}^p$$

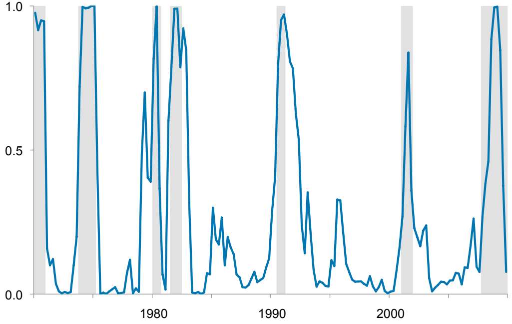
- Want to predict the next data point(s)

$$Y_{n+1}, \ldots, Y_{n+k}$$



Source: Czech National Bank

METHODS OF 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

Bias-variance again

Generate data from the following model:

$$Y_t = .9Y_{t-1} - .81Y_{t-2} + .72Y_{t-3} - \cdots - .1Y_{t-10} + \epsilon_t$$

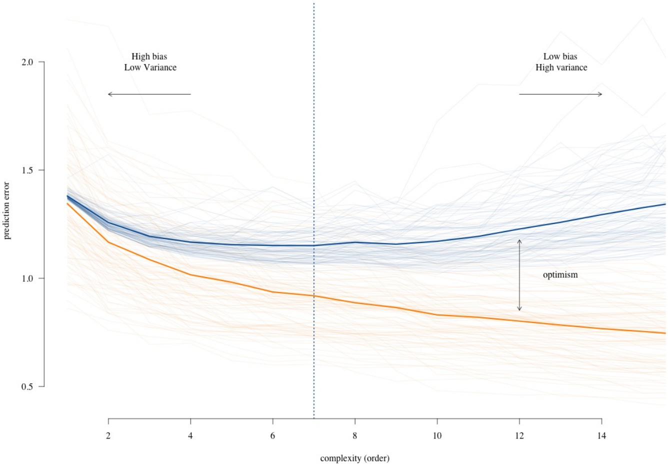
for $t = 1, \ldots, 75$, $\epsilon_t \overset{iid}{\sim} \mathrm{N}(0, 1)$.

This is a stationary AR(10).

Generated 100 realizations of this data and fit AR(1) up to AR(15) to each dataset.

Note that for an AR(k), we are fitting

$$\mathbb{E}[Y_t | (Y_n)_{n<t}] = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_k Y_{t-k}$$


High bias
Low Variance
Low bias
High variance
optimism
prediction error
complexity (order)
2.0
1.5
1.0
0.5
2
4
6
8
10
12
14

Caveats and qualifications

Warning!

- Time series data are hard
- Today you get 5 minutes
- Many omissions / simplifications
- The main points still hold

Disclaimer

I'm going to critique some economic forecasting methods. My comments do not generalize to particular individuals / schools of thought / econometricians generally / etc. I'm just presenting a statisticians' view of some particular practices that I'm suggesting young scholars may want to avoid.

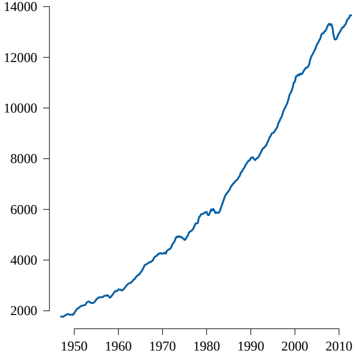
Today

Economic forecasting tools

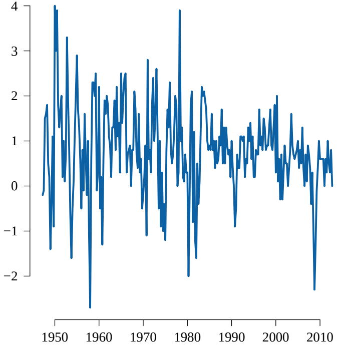
1. Hodrick-Prescott Filter - Data preparation
2. Vector Auto Regressions (Bayesian if you like)
3. DSGEs

STATIONARITY

- Issue 1 – Many common TS methods need “stationary” process



GDP in $ Billions



Quarterly % change

Source: Federal Reserve Economic Database: GDPC1

Inducing stationarity

- Some techniques can handle some types of nonstationarity (Markov switching, general state-space)
- Many cannot (autoregressions)
- Inference is hard in nonstationary settings ("parameters" change with time)
- Economic theory often requires stationary data – "Equilibrium"

How do we make nonstationary data stationary?

Inducing stationarity

Suppose $Y_1, Y_2, \ldots$ is a time series.

First differences:	Under some conditions, the difference $Y_t - Y_{t-1}$ may be stationary (first order integrated)
Detrend:	Suppose you can model Y_t as deviations from some trend, just remove the trend, model the deviations

Test to see if the result is stationary. Tests assume "nice" model of some form

Most of this is not necessary if we just want good predictions: we can predict the weather pretty well, and it's not stationary

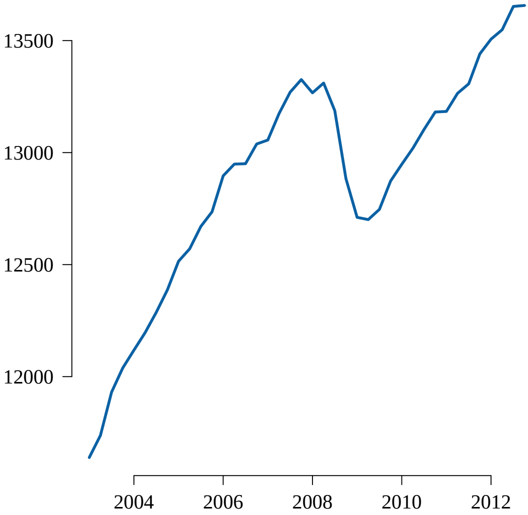
Detrending

Suppose we want to do this anyway.

Remember, economies are just deviations around "equilibrium" . . .

And that "equilibrium" grows smoothly

Oops



US GDP up close

HP Filter

The Hodrick-Prescott Filter (1997).

Estimate a smooth trend $\widehat{g}$.

$$\widehat{g} = \underset{g}{\operatorname{argmin}} ||Y - g||_2^2 + \lambda ||\Delta g||_2^2$$

where $\lambda = 1600$ (based on empirical evaluations for Quarterly GDP in 1997 – does not generalize)

Special case of a "smoothing spline" (see Wahba (1978,1990, others)). Compare to Ridge Regression.

HP FILTER

2 PROBLEMS

1. The estimated trend $\widehat{g}$ is highly sensitive to λ (need to carefully choose tuning parameters)
2. The HP filter is a 2-sided smoother (should be called “HP smoother”)

Consequences:

1. The magnitude of the residuals is completely determined by λ
2. Once you use the HP filter, you cannot evaluate the “forecasts”
3. The HP filter uses the future as well as the past to calculate the trend. Thus residuals may be stationary, but they depend on the future

To `R` $\rightarrow$ for a demonstration!

Now that we’ve detrended, how do we forecast?

Vector autoregressions

Application of (multiple) least squares linear regression to time series:

- Observe a vector $Y_t \in \mathbb{R}^q$ of data at each time $t = 1, \ldots, n$
- Say $Y_t =$ (CPI, GDP, Unemployment, …, Industrial Production)
- Use past Y's to linearly predict future Y's
- Compare to autoregression earlier

Model:

$$Y_t^\top = Y_{t-1}^\top B_1 + Y_{t-2}^\top B_2 + \cdots + Y_{t-p}^\top B_p + E_t$$

where $B_j \in \mathbb{R}^{q \times q}$.

Optimization form

Best linear predictor with memory p

$$\widehat{\mathbb{B}} = \underset{\mathbb{B}}{\text{argmin}} \, |||\mathbb{Y} - \mathbb{X}\mathbb{B}|||_F^2$$

$$\mathbb{Y} = \begin{bmatrix} Y_{p+1}^\top \\ Y_{p+2}^\top \\ \vdots \\ Y_n^\top \end{bmatrix} \qquad \mathbb{X} = \begin{bmatrix} Y_p^\top & Y_{p-1}^\top & \cdots & Y_1^\top \\ Y_{p+1}^\top & Y_p^\top & \cdots & Y_2^\top \\ Y_{p+2}^\top & & & \\ & & \vdots & \\ Y_{n-1}^\top & Y_{n-2}^\top & \cdots & Y_{n-p-1}^\top \end{bmatrix}$$

$$\mathbb{B} = \begin{bmatrix} B_1 & B_2 & \cdots & B_p \end{bmatrix}^\top$$

Problem:
With q time series and p lags, this has $q^2 p$ parameters. We risk overfitting.

Solution: Regularize!

BVARs

Economists have been regularizing VARs for a long time. They call them Bayesian Vector Autoregressions (see Doan, Litterman, and Sims (1984))

Recall that to use ridge regression, all the covariates must be on the same scale. Same is true here. Assume that each time series is normalized to have mean 0 and variance 1.

BVARs

The ridge version of the VAR is

$$\widehat{\mathbb{B}}_{ridge} = \underset{\mathbb{B}}{\operatorname{argmin}} |||\mathbb{Y} - \mathbb{X}\mathbb{B}|||_F^2 + \lambda|||\mathbb{B}|||_F^2.$$

This shrinks all the coefficients toward 0 just like in Ridge Regression. Again, this corresponds to the posterior mean in a Bayesian setting with prior $B_{ij} \overset{iid}{\sim} \mathrm{N}(0, \lambda^{-1/2})$

And it has the same closed form solution:

$$\widehat{\mathbb{B}}_{ridge} = (\mathbb{X}^\top \mathbb{X} + \lambda \mathbb{I})^{-1} \mathbb{X}^\top \mathbb{Y}$$

BVARs: Minnesota prior

Generally in economics, the prior is altered so that own lags are shrunk toward 1 while cross lags are shrunk toward 0 (the Minnesota Prior)

Also, some decay is usually enforced so that the coefficients decay in the lag (I'm going to ignore this here)

The optimization problem then becomes:

$$\widehat{\mathbb{B}}_M = \underset{\mathbb{B}}{\operatorname{argmin}} |||\mathbb{Y} - \mathbb{X}\mathbb{B}|||_F^2 + \lambda|||\mathbb{B} - \mathbb{G}|||_F^2,$$

where $\mathbb{G} = \begin{bmatrix} \mathbb{I}_k & \mathbb{I}_k & \cdots & \mathbb{I}_k \end{bmatrix}$.

λ is often called the "overall tightness" and must be carefully chosen.

$$\widehat{\mathbb{B}}_M = (\mathbb{X}^\top \mathbb{X} + \lambda \mathbb{I})^{-1} (\mathbb{X}^\top \mathbb{Y} - \lambda \mathbb{G})$$

BVAR EXAMPLE

Generate data from the following VAR:

$$\begin{bmatrix} Y_{1,t} & Y_{2,t} \end{bmatrix} = \begin{bmatrix} Y_{1,t-1} & Y_{2,t-1} \end{bmatrix} \begin{bmatrix} 0.9 & 0 \\ 0 & 0.9 \end{bmatrix} + E_t$$

where $E_{1,t}$ and $E_{2,t}$ are independent standard normal random variables.

Fit a BVAR with Minnesota prior

To `R` $\rightarrow$ for a demonstration!

You must pick the overall tightness carefully!

Location, location, location

Notice that the Minnesota prior is wrong in my example.

True	Belief
$\mathbb{B} = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.9 \end{bmatrix}$	$\mathbb{B} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

That's why forecasts are so awful when λ gets large. $\Rightarrow$ "tight" prior

This is true even when the model is correct! I really did get data from a VAR, just not the one the Minnesota prior thinks I did.

Things are much worse if I put a prior on an incorrect model. Especially a tight prior.

Large n can wipe out the prior for fixed λ.

What if I want to use everything I know about microeconomic behavior?

Rationality and utility maximization and such?

The state of the art

I want to forecast the macroeconomy.

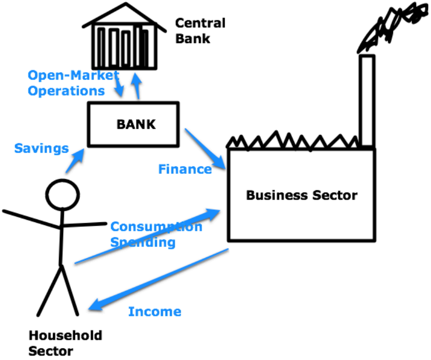
I want my forecasts to derive from economic theory.

I use a Dynamic Stochastic General Equilibrium model.

Are my forecasts any good? Can I make them any better? Can I make inferences about behavior based on my fitted DSGE model?

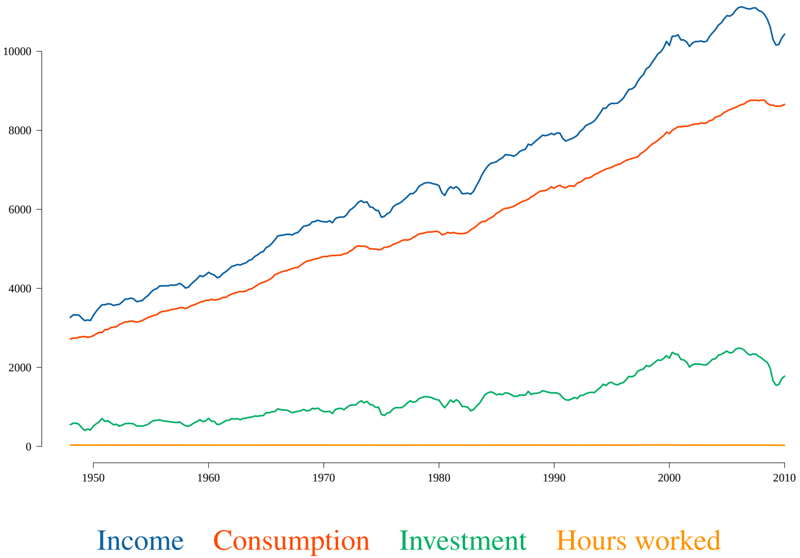
DSGE MODELS

- 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
- Have come under fire for being unable to forecast the financial collapse of 2008–?



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

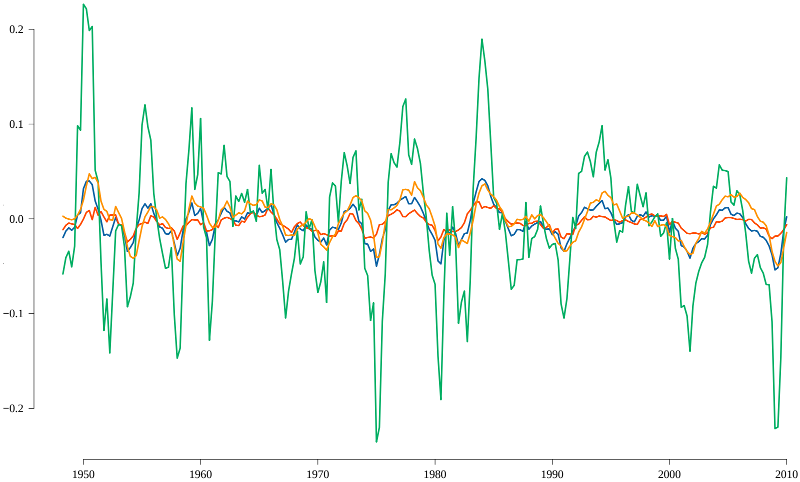
DATA (1948:I–2011:IV)



Cleaning the data

- DSGE models deal with deviations about "equilibrium"
- To get these deviations, I have to find the assumed "equilibrium"
- Like most, I detrend using the HP filter
- I used a one-sided version so that forecasts make sense
- I did not choose the tuning parameter
- Relative to a linear trend, the filter removes $> 99\%$ of the variability of GDP around the trend
- DSGE doesn't predict trend. Uh oh…

Some data (1948:I–2011:IV)



Income Consumption Investment Hours worked

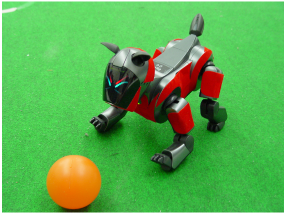
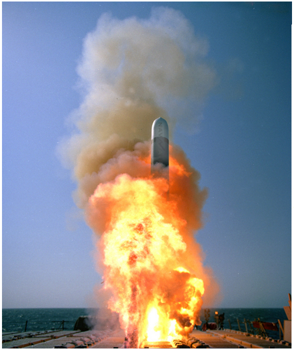
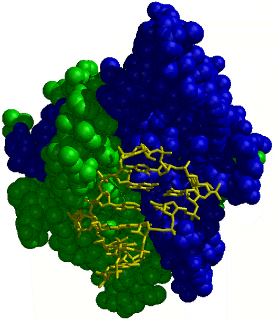
RBC MODEL

- Imagine an infinitely-long-lived individual who faces the following constrained optimization problem:

$$
\begin{aligned}
\max_{c_t, l_t} U &= \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t, l_t) \\
y_t &= z_t q(k_t, n_t) \\
1 &= n_t + l_t \\
y_t &= c_t + i_t \\
k_{t+1} &= i_t + (1-\delta) k_t \\
z_t &\sim \mathrm{AR}(1)
\end{aligned}
$$

State space models detour

- Lots of disciplines use state space models
- Sometimes motivated directly by physical relationships

State space models detour

State equation:

$$\alpha_{t+1} = T\alpha_t + \eta_{t+1}$$

Observation equation:

$$y_t = A\alpha_t + \epsilon_t$$

Some assumptions:

$$0 = \mathbb{E}[\eta_t] = \mathbb{E}[\epsilon_t]$$

I get to observe y_t, I never see α_t "hidden state"

Relationship to state space models

DSGE Model

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

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

$$1 = n_t + l_t$$

$$y_t = c_t + i_t$$

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

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

Math games

State Space Model

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

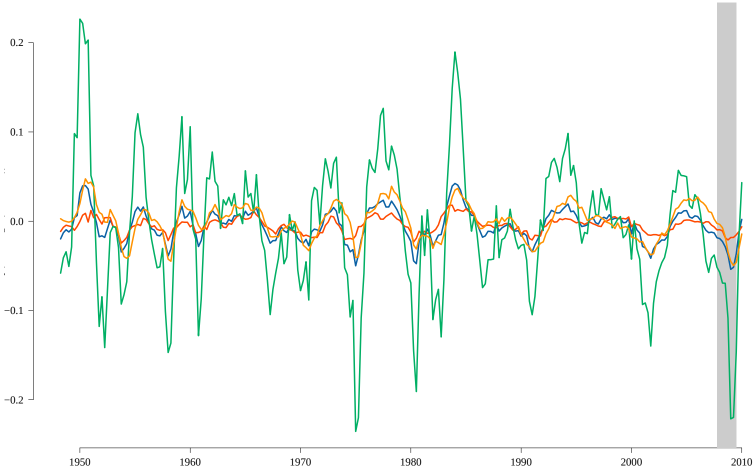
$$\boldsymbol{\alpha}_{t+1} = h(\boldsymbol{\alpha}_t, \eta_{t+1})$$

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

DSGE AS PRIOR

- The DSGE works as a prior on the underlying state-space model
- The quality of the forecasts depends on the tightness and location of the regularization
- Altering the DSGE (adding restrictions, assumptions, sticky prices, etc) changes both
- IF this makes the DSGE more realistic, it corresponds to decreasing bias and increasing variance
- It is not clear that this is a good idea.

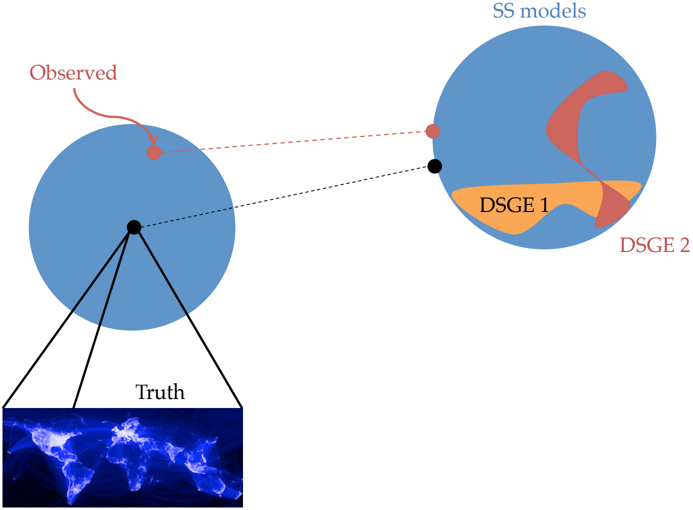
THE PROBLEM



Income Consumption Investment Hours worked

MSE: Non-recession .89 During recession 2.11

The world economy. According to Facebook.


SS models
Observed
DSGE 1
DSGE 2
Truth

What went wrong?

The issues

- DSGEs do not and cannot describe the real economy
- Neither can state-space models, VARs, or anything else
- We might still be able to predict well
- Why didn't DSGEs at least predict well?

WHY DID IT FAIL?

- Those were residuals after detrending.
- The HP filter removed the “trend”. That big recession? HP filter thought that was “trend”. Can’t forecast what’s not in the data (especially with a 2-sided filter).
- It missed the residuals too, but that was $< 1\%$ of the downturn.
- The US Congress held hearings on this issue.

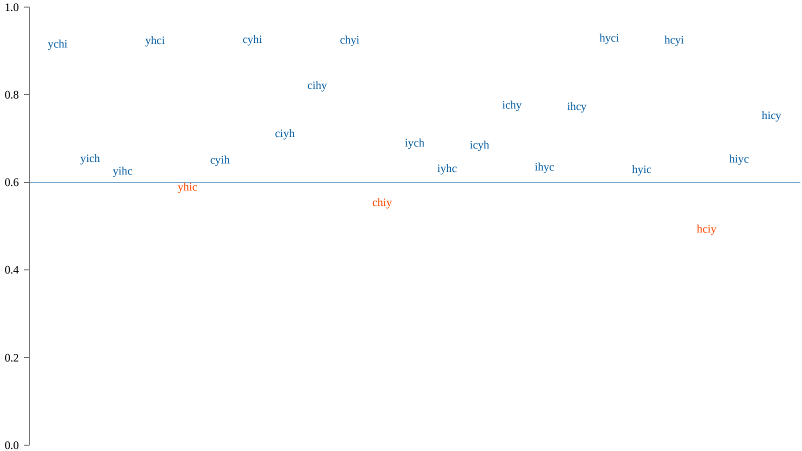
Are DSGEs good for forecasting?

- Requires separate model for the trend
- Puts strange nonlinear constraints on the space of models
- Don’t know if these constraints help or hurt: the suspicion is hurt (next slide)
- Any state space model is wrong. A DSGE is a SS model. Therefore it is wrong.
- Can it forecast better than an unconstrained SS model?
- What determines its forecasting abilities?

Do DSGEs really impose "economically relevant" constraints?

1. The point of a DSGE is that it adds economic theory that is otherwise absent from the statistical model
2. If this model fits the data well, then the DSGE is a good approximation to the real economy
3. What does it mean to "fit the data" without overfitting?
4. Try an experiment: mess up the data, see if the DSGE notices
5. Mislabel the time series. Call Income $\rightarrow$ Consumption, etc.

% MSE remaining (relative to predicting w/ 0)



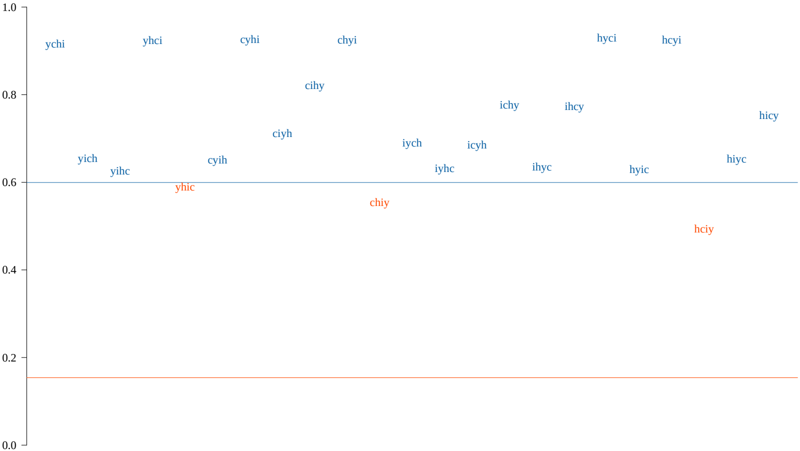
Horizontal line: DSGE w/ real data

I don’t know any economic theory.
Can I beat the DSGE?

The alternative

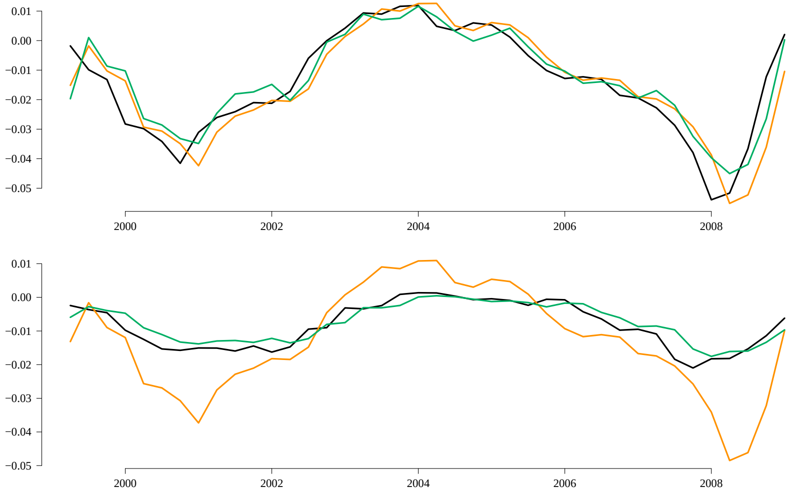
- I used a VAR(4).
- I combined this with a lasso penalty to avoid overfitting
- I choose the tuning parameter with cross validation
- The model knows nothing about recent economic events
- It took about 10 lines of code and less than 1 second to run
- The DSGE takes about 250 lines of code and 30 minutes to run

% MSE remaining



DSGE w/ real data   Alternative

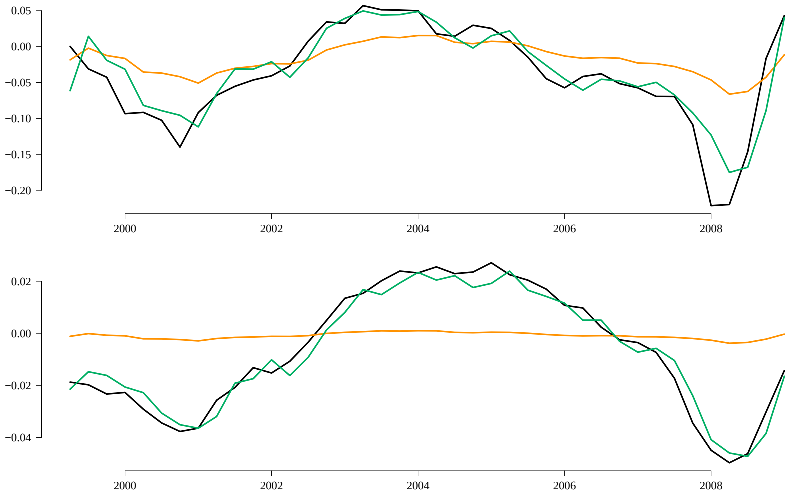
Breaking it down



Top – Income/Output   Bottom – Consumption

DSGE   Alternative

Breaking it down



Top – Investment   Bottom – Hours Worked

DSGE   Alternative

Bringing it all together

- DSGEs cannot be used to make forecasts when using the HP filter
- You CAN use a 1-sided filter instead
- The filter explains most of the variation in macroeconomic time-series
- Very little is left for the DSGE
- The DSGE is not terrible. It explains about 40% of remaining variation
- The DSGE imposes strange priors on the state space model which are not really economically relevant

The moral

Carefully constructed theoretical models are good.

If they match the data well (without overfitting), we can imagine using them for predictions and inference.

Priors and regularization are good. They help us avoid overfitting.

But bad priors can give bad results.

When the prior can't distinguish the data, something is wrong.

Simple models with carefully chosen tuning parameters work better.

Up next:
How to predict recessions and other classification tricks