Likelihoods and Bayes

(See Statistics)

Trevor Gallen

Example

- Take a two-state Markov chain, in which you observe the following data:

$$X_t = \{2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 1\}$$

- Say we wanted to estimate the markov process:

$$\pi = \begin{bmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{bmatrix}$$

- How do we do it? (What is your estimate?)

Maximum Likelihood

- We can write down the likelihood of seeing each type of transition:

$$\mathcal{L}_{1\to1} = p_{11}$$
$$\mathcal{L}_{1\to2} = p_{12}$$
$$\mathcal{L}_{2\to1} = p_{21}$$
$$\mathcal{L}_{2\to2} = p_{22}$$

Or:

$$\mathcal{L} = \prod_{i=1}^{T} (\mathcal{L}_{1\to1})^{x_{11}} (\mathcal{L}_{1\to2})^{x_{12}} (\mathcal{L}_{2\to1})^{x_{21}} (\mathcal{L}_{2\to2})^{x_{22}}$$

Maximum Likelihood

- The likelihood:

$$\mathcal{L} = \prod_{i=1}^{T} (p_{11})^{x_{11}} (1 - p_{11})^{x_{12}} (1 - p_{22})^{x_{21}} (p_{22})^{x_{22}}$$

- Taking logs:

$$\log \mathcal{L} = \sum_{i=1}^{T} x_{11} \log(p_{11}) + x_{12} \log(1 - p_{11}) + x_{21} \log(1 - p_{22}) + x_{22} \log(p_{22})$$

- Is this all we can do?

Maximum Likelihood

- The likelihood:

$$\mathcal{L} = \prod_{i=1}^{T} (p_{11})^{x_{11}} (1 - p_{11})^{x_{12}} (1 - p_{22})^{x_{21}} (p_{22})^{x_{22}}$$

- Taking logs:

$$\log \mathcal{L} = \sum_{i=1}^{T} x_{11} \log(p_{11}) + x_{12} \log(1 - p_{11}) \\ + x_{21} \log(1 - p_{22}) + x_{22} \log(p_{22})$$

- Is this all we can do? (Hint: no.)

Maximum Likelihood

- The likelihood:

$$\mathcal{L} = \prod_{i=1}^{T} (p_{11})^{x_{11}} (1 - p_{11})^{x_{12}} (1 - p_{22})^{x_{21}} (p_{22})^{x_{22}}$$

- Taking logs:

$$\begin{aligned} \log \mathcal{L} = \sum_{i=1}^{T} & x_{11} \log(p_{11}) + x_{12} \log(1 - p_{11}) \\ & + x_{21} \log(1 - p_{22}) + x_{22} \log(p_{22}) \end{aligned}$$

- Is this all we can do? (Hint: no.)

- The first observation gives us data!

Better Maximum Likelihood

- The likelihood:

$$\log \mathcal{L} = \sum_{i=1}^{T} x_{11} \log(p_{11}) + x_{12} \log(1 - p_{11}) \\ + x_{21} \log(1 - p_{22}) + x_{22} \log(p_{22})$$

- Denote a dummy for the first observation as $x^{[1]}$ or $x^{[2]}$, depending on the value:

$$\log \mathcal{L} = \sum_{i=1}^{T} x_{11} \log(p_{11}) + x_{12} \log(1 - p_{11}) \\ + x_{21} \log(1 - p_{22}) + x_{22} \log(p_{22}) \\ + x^{[1]} \left(\frac{1 - p_{22}}{2 - p_{11} - p_{22}} \right) + x^{[2]} \left(\frac{1 - p_{11}}{2 - p_{11} - p_{22}} \right)$$

- Why?

Results

- We could do things closed form (how?)
- Or numerically (see Markov.m)
- What else can we do?

Bayes Rule

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

More Fun with Markovs

- Now imagine we didn't observe the states x_t directly, but observed some noise process y_t
- If state $x_t = 1$, then $y_t \sim \mathcal{N}(\mu_1, \sigma_1^2)$
- If state $x_t = 2$, then $y_t \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- To skip annoying notation for the first step, let's say we knew the first state but from then on out we knew nothing else.

Example

$$x_1 = 1$$

$$\pi = \begin{bmatrix} 0.9 & 0.1 \\ 0.95 & 0.05 \end{bmatrix}$$

$$y_t | x_t = 1 \sim \mathcal{N}(0, 10)$$

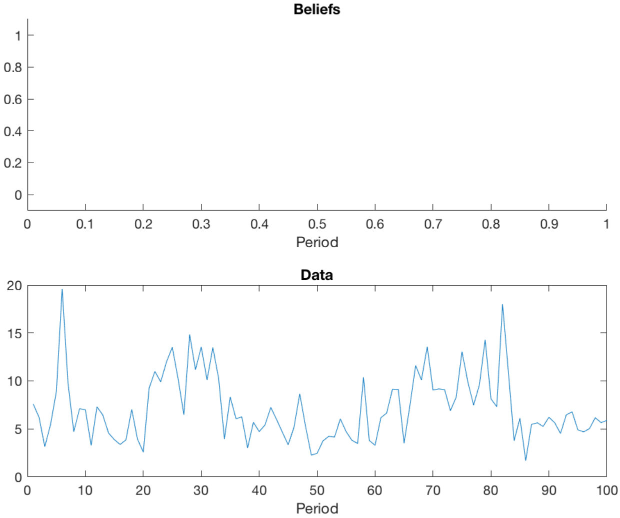
$$y_t | x_t = 2 \sim \mathcal{N}(1, 4)$$

$$P(x_t | y_t) = \frac{P(x_t) P(y_t | x_t)}{P(y_t)}$$

These types of models are called “Regime Switching” models

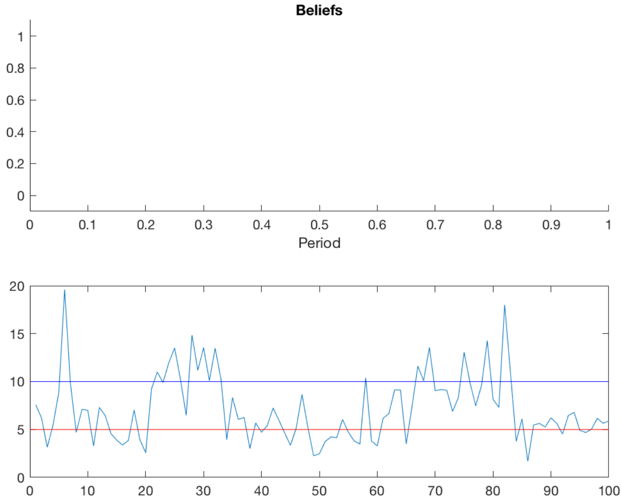
See Markov.2

Example



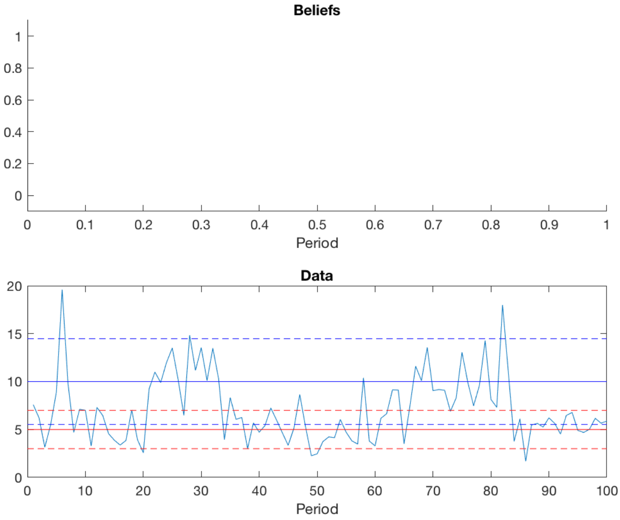
When are we in which state? $y_t|x_t = 1 \sim \mathcal{N}(0, 10)$, $y_t|x_t = 2 \sim \mathcal{N}(1, 4)$

Example



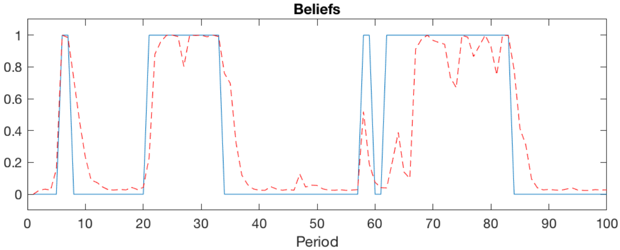
Graphing the means is relevant

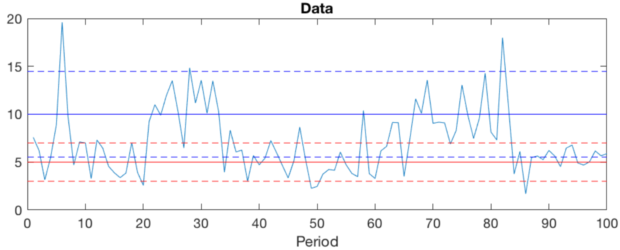
Example



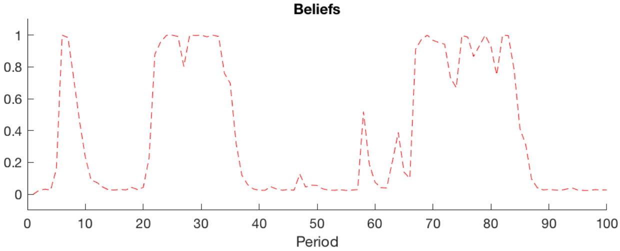
Graphing the 2sd interval is relevant

Example

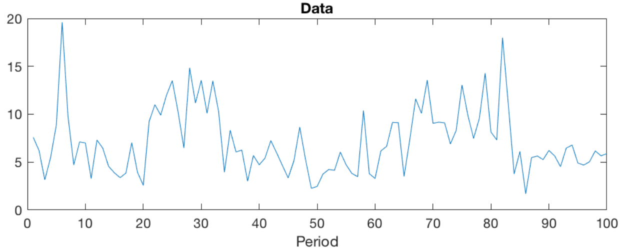

Beliefs
1
0.8
0.6
0.4
0.2
0
0 10 20 30 40 50 60 70 80 90 100
Period


Data
20
15
10
5
0
0 10 20 30 40 50 60 70 80 90 100
Period

Example

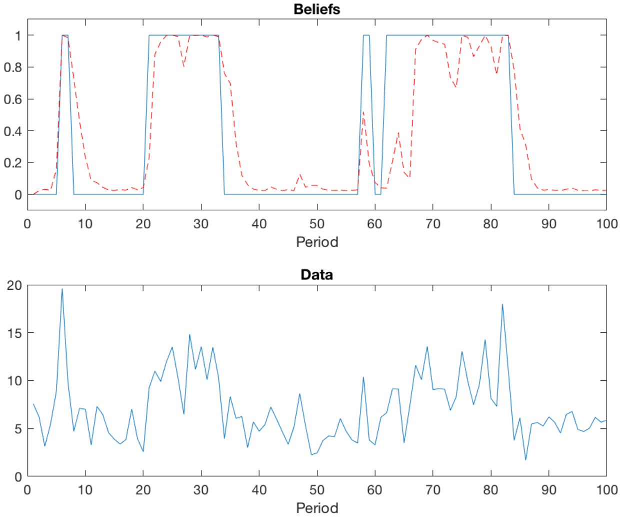


Beliefs
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
Period



Data
20
15
10
5
0
0
10
20
30
40
50
60
70
80
90
100
Period

Example



Beliefs
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
Period
Data
20
15
10
5
0
0
10
20
30
40
50
60
70
80
90
100
Period

Kalman Filter: Motivation

- Life is full of scenarios in which we see a *signal* about some true underlying process but never observe the truth
  - Missiles
  - Polls
  - Recessions
  - Economic variables
- We typically have some belief of what the underlying object is, where it's going to go, and get some signal related to the object
- How do we put all our information together?

Kalman Filter: Preview to Lemma

Let X explain Y:

$$Y = X\beta + \epsilon$$

Then:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

And:

$$\hat{Y} = X\beta$$

So:

$$\begin{aligned} Var(Y - \hat{Y}|X) &= Var(Y - X\beta|X) \\ &= Var(Y|X) + \beta^2 \underbrace{Var(X|X)}_{0} - 2\beta Cov(Y, X|X)) \\ &= Var(Y|X) - 2\beta \frac{\beta}{Var(X)} \\ &= Var(Y|X) + 2\beta^2 Var(X)^{-1} \end{aligned}$$

Kalman Filter: Lemma*

If:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} S_{XX'} & S_{XY'} \\ S_{YX'} & S_{YY'} \end{bmatrix} \right)$$

Then:

$$Y|X \sim \mathcal{N}\left(S_{XY'} S_{XX'}^{-1} X, S_{YY'|X} \right)$$

Where, letting $A = S_{XY'} S_{XX}^{-1}$

$$S_{YY'|X} = S_{YY'} - S_{XY'} S_{XX}^{-1} S_{YX'} = S_{YY'} - A S_{XX'}^{-1} A'$$

Matrix analogue of previous slide:

$$Var(Y - \hat{Y}|X) = Var(Y|X) + 2\beta^2 Var(X)^{-1}$$

In other words, our expectation of X given Y comes from a regression, and our conditional variance is our unconditional variance minus the regression coefficient squared times the variance of our signal.

* This and the next two slides are inspired by Harald Uhlig's notation & wonderful slides.

Kalman Motivation

- What if you observe some **possibly noisy** vector of values Y and want to use them to track some **possibly moving** vector of values ξ?
  - Bayesian Estimation: Data and Coefficents
  - Polling: Poll and Latent Preferences
  - Permanent Income: period income and consumption and Agent beliefs about permanent income
  - Ballistic missiles, cars: noisy GPS signal and actual location
- Want to combine noisy observations, unobservable moving due to some law of motion (with noise)
- This is the Kalman Filter

Intuition/Shape of Kalman Algorithm

- Start with beliefs (mean, variance) about ξ this period (not conditional on information this period)
- We therefore have beliefs about what Y will be (both itself noisy + don't know ξ in truth)
- Observe Y
- Degree to which we are surprised by Y, plus how much we trust it, determines how we change our beliefs about mean and variance of ξ (conditional on information this period)
- Now we have (mean, variance) beliefs about today conditional on data today
- Forecast (mean, variance) beliefs about tomorrow using law of motion
- Repeat

Kalman System

- You have an observation equation:

$$Y_t = H_t \xi_t + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \Sigma_t)$$

- And a state equation:

$$\xi_{t+1} = F_{t+1} \xi_t + \eta_{t+1} \qquad \eta_{t+1} \sim \mathcal{N}(0, \Phi_{t+1})$$

- We assume that ϵ_t and η_t are independent.
- Y_t is a noisy observation of ξ_t, which moves around with noise.

Updating our beliefs

- We start with some beliefs from last period about where ξ would be this period (called $\xi_{t|t-1}$).
- We summarize these as:

$$\xi_{t|t-1} \sim \mathcal{N}\left(\hat{\xi}_{t|t-1}, \Omega_{t|t-1}\right)$$

- We want to look at information today and say what we think ξ is, calling this $\xi_{t|t}$.

First Step: best guess of what the signal will be

- We start with beliefs:

$$\xi_t \sim \mathcal{N}\left(\hat{\xi}_{t|t-1}, \Omega_{t|t-1}\right)$$

- And we know, as a law:

$$Y_t = H_t \xi_t + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \Sigma_t)$$

- Then we have our best guess of what Y will be, along with its variance:

$$\hat{Y}_t = H_t \xi_{t|t-1}$$

$$S_{YY|t} = H_t \Omega_{t|t-1} H_t + \Sigma_t$$

Second Step: use surprise info to update prior beliefs

- We have our *unexpected* information:

$$\hat{\epsilon}_t = Y_t - \hat{Y}_t$$

- Then our best fit is, like a regression fit:

$$\hat{\xi}_{t|t} = \hat{\xi}_{t|t-1} + S_{\xi Y'|t} S^{-1}_{YY'|t} \hat{\epsilon}_t$$

- Where our "signal" is:

$$S_{\xi Y'|t} = \Omega_{t|t-1} H'_t$$

- And our beliefs are updated:

$$\Omega_{t|t} = \Omega_{t|t-1} - S_{\xi Y'|t} S^{-1}_{YY'|t} S_{Y\xi'|t}$$

Third Step: use current best beliefs to find tomorrow's best beliefs

- We have our beliefs for today, $\hat{\xi}_{t|t}$ and $\Omega_{t|t}$
- We want:

$$\xi_{t+1} \sim \mathcal{N}(\hat{\xi}_{t+1|t}, \Omega_{t+1|t})$$

- Update using the law of motion:

$$\hat{\xi}_{t+1|t} = F_{t+1}\hat{\xi}_{t|t}$$

$$\Omega_{t+1|t} = F_{t+1}\Omega_{t|t}F'_{t+1} + \Phi_{t+1}$$

Summarizing the Kalman Filter

$$Y_t \sim \mathcal{N}\left(H_t \xi_t, \Sigma_t\right), \qquad \xi_t \sim \mathcal{N}\left(F_{t+1}\xi_t, \Phi_{t+1}\right)$$

- Given $\xi_t \sim \mathcal{N}(\hat{\xi}_{t|t-1}, \Omega_{t|t-1})$,
  1. Forecast Y_t given what you know:

$$\hat{Y}_t = H_t \hat{\xi}_{t|t-1} \qquad S_{YY'|t} = H_t \Omega_{t|t-1} H_t' + \Sigma_t$$

  2. Update ξ_t given surprise:

$$\hat{\xi}_{t|t} = \hat{\xi}_{t|t-1} + S_{\xi Y'|t} S^{-1}_{YY'|t}(Y_t - \hat{Y}_t)$$

$$\hat{\Omega}_{t|t} = \hat{\Omega}_{t|t-1} + S_{\xi Y'|t} S^{-1}_{YY'|t} S_{\xi Y'|t}$$

     Where: $S_{\xi Y'|t} = \Omega_{t|t-1} H_t'$
  3. Forecast and set up for tomorrow

$$\hat{\xi}_{t+1|t} = F_{t+1}\hat{\xi}_{t|t} \qquad \Omega_{t+t|t} = F_{t+1}\Omega_{t|t}F_{t+1}' + \Phi_{t+1}$$

Coding it

See Kalman.m

Uses

- Estimating underlying data, like polls, recessions
- Alternatively, think of your *regression coefficients* as your unknown ξ and your data as Y_t
- Then for:

$$Y_t \sim \mathcal{N}\left(H_t \xi_t, \Sigma_t\right), \qquad \xi_t \sim \mathcal{N}\left(F_{t+1}\xi_t, \Phi_{t+1}\right)$$

  - Y_t is your dependent variable
  - H_t is your independent variable
  - Σ_t is your noise term
  - F_{t+1} is just 1, if your coefficients are constant
  - Φ_{t+1} is zero, if your coefficients are constant.
- Now you can run Kalman filter point-by-point on your data to uncover your belief *distribution* over your coefficients.

Kalman Smoother

- Note that your beliefs at any given point are optimal given initial beliefs conditional on information *up to that point*
- But beliefs about today weren't conditioned on tomorrow's data
- But tomorrow tells us something about today!
- Once run Kalman Filter, you have beliefs about ξ conditional on all past data, can run in reverse, condition beliefs about yesterday on today's data
- This is the Kalman Smoother

One More Example: BVars (DSGE estimation)

- Standard feel of a DSGE estimation: **parameters are unknown state**, and *data take place of parameters*

$$\underbrace{\begin{bmatrix} c_t \\ y_t \\ i_t \end{bmatrix}}_{Y_t} = \underbrace{\overbrace{\begin{bmatrix} y_{t-1} & 0 & 0 \\ 0 & y_{t-1} & 0 \\ 0 & 0 & y_{t-1} \end{bmatrix}}^{\text{Data!}}}_{H_t} \underbrace{\overbrace{\begin{bmatrix} \beta_{1,t} \\ \beta_{2,t} \\ \beta_{3,t} \end{bmatrix}}^{\text{Parameters}}}_{\xi_t} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \epsilon_{3,t} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma^2_{\epsilon_1} & \cdot & \cdot \\ \cdot & \sigma^2_{\epsilon_2} & \cdot \\ \cdot & \cdot & \sigma^2_{\epsilon_3} \end{bmatrix}$$

And the "state equation" is:

$$\begin{bmatrix} \beta_{1,t+1} \\ \beta_{2,t+1} \\ \beta_{3,t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{F} \begin{bmatrix} \beta_{1,t} \\ \beta_{2,t} \\ \beta_{3,t} \end{bmatrix} \qquad \Phi = \begin{bmatrix} 0 \end{bmatrix}$$

Our use: Insurance

- As labor, macro, and public economists, we care about people's lifecycle income paths
- We know a random walk component is important
- We know transitory shocks are also important
- Write down a simple flexible model of income $y_{i,t}$ as the sum of permanent income $z_{i,t}$ and perfectly transitory shock $\epsilon_{i,t}$.

$$y_{i,t} = z_{i,t} + \epsilon_{i,t}$$

- Where permanent income is subject to a shock $\zeta_{i,t}$

$$z_{i,t} = z_{i,t-1} + \zeta_{i,t}$$

- Note: for more about this, read Meghir and Pistaferri (ECMA 2004) or Blundell, Pistiaferri, and Preston (AER 2008)

Our use: Insurance

- We can therefore write the *change* in income as:

$$y_{i,t} = z_{i,t} + \epsilon_{i,t}$$

- If we're interested in what consumption tells us about income, we might also have:

$$c_{i,t} = \phi z_{i,t} + \psi \epsilon_{i,t} + \xi_{i,t}$$

- Where ϕ and ψ are the MPC out of permanent and temporary income changes, respectively. $\xi_{i,t}$ denotes shocks that change consumption, but not income (such as lifecycle concerns, so it may be common between individuals).
- This allows both y and c to tell us about permanent income
- Two observations: income and consumption
- Two (or three) hidden variables: "true" permanent income $z_{i,t}$, transitory income $\epsilon_{i,t}$ and transitory consumption shock $\xi_{i,t}$.

Aside: What's the motivation?

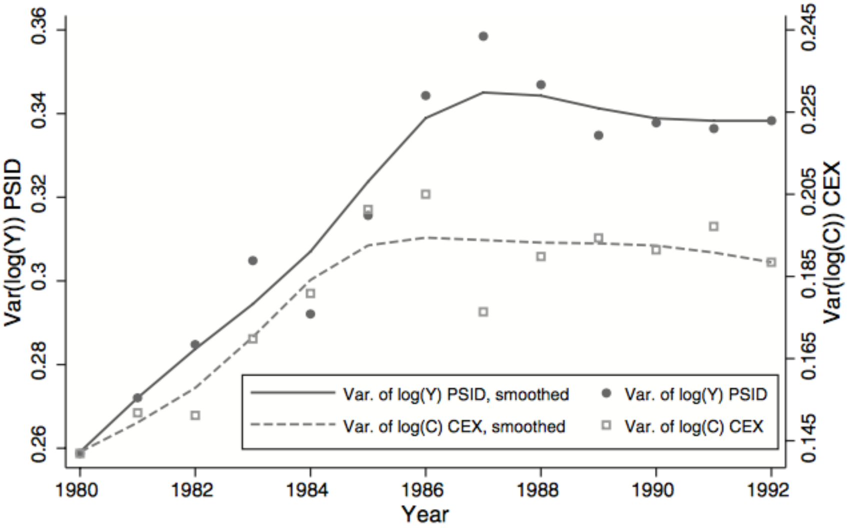
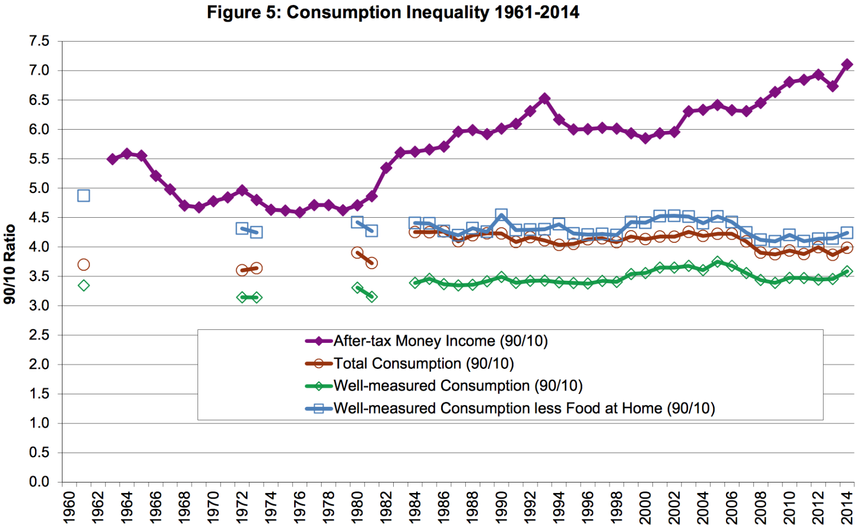


Figure 1. Overall Pattern of Inequality

Aside: Consumption inequality



Applying the Kalman Filter's notation for H, F, Σ, Φ, ξ, and Y, above, we can write out the "observation equation":

$$\underbrace{\begin{bmatrix} c_{i,t} \\ y_{i,t} \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} \phi \\ 1 \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} z_{i,t} \end{bmatrix}}_{\xi_t} + \begin{bmatrix} \psi\epsilon_{i,t} + \xi_{i,t} \\ \epsilon_{i,t} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \psi^2\sigma_\epsilon^2 + \sigma_\xi^2 & \psi \\ \psi & \sigma^2\epsilon \end{bmatrix}$$

And the "state equation" is:

$$\begin{bmatrix} z_{i,t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \end{bmatrix}}_{F} \begin{bmatrix} z_{i,t} \end{bmatrix}$$

$$\Phi = \begin{bmatrix} \sigma_\zeta^2 \end{bmatrix}$$

Identification

- But how do we get the parameter values out? The variances?
- Can estimate from conditional moments
- For instance:

$$
\begin{aligned}
E(\Delta y_t \Delta y_{t-1}) &= E((\zeta_{i,t} + \Delta\epsilon_{i,t})(\zeta_{i,t-1} + \Delta\epsilon_{i,t-1})) \\
&= E(\zeta_{i,t}\zeta_{i,t-1} + \Delta\epsilon_{i,t}\zeta_{i,t-1} + \zeta_{i,t}\Delta\epsilon_{i,t-1} + \Delta\epsilon_{i,t}\Delta\epsilon_{i,t-1}) \\
&= E((\epsilon_{i,t} - \epsilon_{i,t-1})(\epsilon_{i,t-1} - \epsilon_{i,t-2})) \\
&= E(-\epsilon_{i,t-1}\epsilon_{i,t-1}) \\
&= -\sigma_\epsilon^2
\end{aligned}
$$

- Intuition: the only reason the change in income today & change in income yesterday are correlated is the transitory shock, which shows up in both periods (one coming in, the other going out).

Our use

- Similiarly:

$$
\begin{aligned}
E(\Delta c_t \Delta c_{t-1}) &= E\left((\phi\zeta_{i,t} + \psi\Delta\epsilon_{i,t} + \Delta\xi_{i,t})\left(\phi\zeta_{i,t-1} + \psi\Delta\epsilon_{i,t-1} + \Delta\xi_{i,t-1}\right)\right) \\
&= E\left((\psi\Delta\epsilon_{i,t} + \Delta\xi_{i,t})\left(\psi\Delta\epsilon_{i,t-1} + \Delta\xi_{i,t-1}\right)\right) \\
&= E((\psi(\epsilon_{i,t} - \epsilon_{i,t-1}) + (\xi_{i,t} - \xi_{i,t-1}))\left(\psi(\epsilon_{i,t-1} - \epsilon_{i,t-2}) + \right. \\
&\left. + (\xi_{i,t-1} - \xi_{i,t-2})\right)) \\
&= E\left((-\psi\epsilon_{i,t-1} - \xi_{i,t-1})\left(\psi(\epsilon_{i,t-1} - \epsilon_{i,t-2}) + (\xi_{i,t-1} - \xi_{i,t-2})\right)\right) \\
&= E\left(-\psi^2\epsilon_{i,t-1}\epsilon_{i,t-1} - \xi_{i,t-1}\xi_{i,t-1}\right) \\
&= -\psi^2\sigma_\epsilon^2 - \sigma_\xi^2
\end{aligned}
$$

- There are two reasons consumption today and yesterday are correlated. First, the transitory shock (weighted by ψ) increases consumption in the first period relative to the second, in which it falls. Second, consumption has its own "transitory" shock, which comes in just as ϵ did.

Our use

The identifying equations derived above are summarized in Table 1:

Data		Structural Equation
$E(\Delta y_{i,t}\left(\Delta y_{i,t-1} + \Delta y_{i,t} + \Delta y_{i,t+1}\right))$	$=$	σ_ζ^2
$E(\Delta y_t \Delta y_{t-1})$	$=$	$-\sigma_\epsilon^2$
$E(\Delta c_t \Delta c_{t-1})$	$=$	$-\psi^2\sigma_\epsilon^2 - \sigma_\xi^2$
$E(\Delta c_t \Delta y_t)$	$=$	$\phi\sigma_\zeta^2 + 2\psi\Delta\sigma_\epsilon^2$
$E(\Delta c_t(\Delta y_{t-1} + \Delta y_t + \Delta y_{t+1}))$	$=$	$\phi\sigma_\zeta^2$

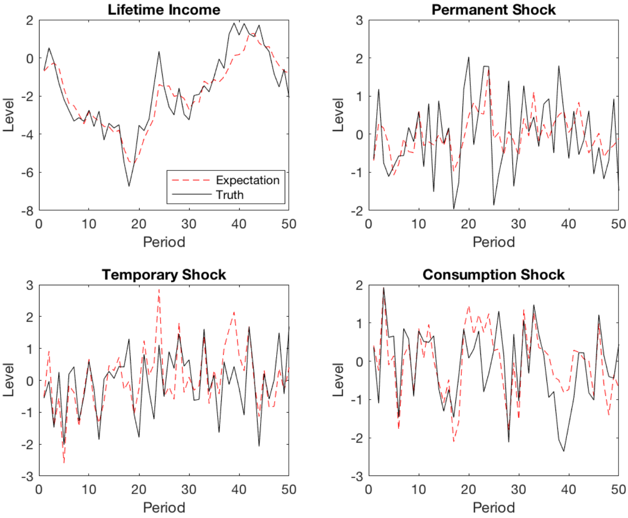
Table: This table summarizes the identifying equations for σ_ζ^2, σ_ξ^2, σ_ϵ^2, ϕ, and ψ.

How good is our estimator?

Description	Parm.	Value	Est. N=10	Est. N=50	Est. N=10000
c to a perm. y	ϕ	1	1.11	0.54	1.02
c to a trans. y	ψ	0.1	0.33	0.09	0.10
StdDev of trans. y	σ_ϵ	0.1	-0.04	0.10	0.10
StdDev of perm. y	σ_ζ	0.1	0.07	0.06	0.10
StdDev of trans. c	σ_ξ	0.1	0.05	0.08	0.11

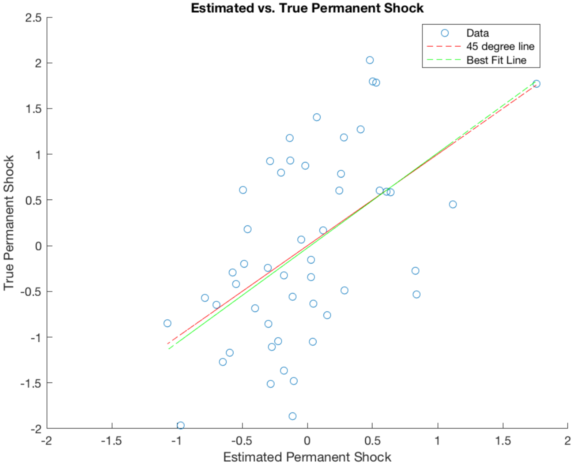
How good is our estimator?

- Once we have the parameters, we can let the Kalman filter rip! Uncover hidden states.



How good is our estimator?

- Compare belief about permanent shock to permanent shock



- $R^2 \approx 0.54$

How good is our estimator?

- Remember, these all have standard errors!

- Can run regressions on these, know how to correct for attrition bias!

Concluding Thoughts

- If you ever do structural estimation with different, hidden regimes, need to filter
- If you ever want to (partially) uncover hidden state (such as permanent income) Kalman Filter
  - Note: might not be super practical in example we gave!
- Very useful in time series estimation
- Used to estimate most macro models (DSGE)