Rational Inattention

November 25, 2016

Motivation

Introspection:

- We know we don't respond to all information available to us without significant economic cost.

Motivation

Introspection:

- We know we don't respond to all information available to us without significant economic cost.

- The price you are willing to pay for a sandwich at lunch should depend — at least a tiny bit — on what has happened to the short term interest rate this morning.

Motivation

Introspection:

- We know we don't respond to all information available to us without significant economic cost.

- The price you are willing to pay for a sandwich at lunch should depend — at least a tiny bit — on what has happened to the short term interest rate this morning.

  But almost no one makes this connection.

Motivation

Anecdotal Observation:

- This might be in part because because the price posted for each sandwich usually is the same from day to day, and when it does change, changes in a discrete jump.

Motivation

Anecdotal Observation:

- This might be in part because because the price posted for each sandwich usually is the same from day to day, and when it does change, changes in a discrete jump.

- But these prices should be varying — at least a tiny bit — with changes in the interest rate, the futures price of beef, wheat and corn, and lots of other things.

Motivation

Anecdotal Observation:

- This might be in part because because the price posted for each sandwich usually is the same from day to day, and when it does change, changes in a discrete jump.

- But these prices should be varying — at least a tiny bit — with changes in the interest rate, the futures price of beef, wheat and corn, and lots of other things.

- Is this because customers have a hard time dealing with changing prices? Because they would not react to small changes? Because the restaurant owner does not find it worthwhile to keep track of and respond to every source of changes in cost at every moment?

Motivation

Macroeconomics:

- Theorists complain that the Smets-Wouters model, and its descendants and ancestor (Christiano, Eichenbaum and Evans), are too full of "ad hoc" sources of inertia.

Motivation

Macroeconomics:

- Theorists complain that the Smets-Wouters model, and its descendants and ancestor (Christiano, Eichenbaum and Evans), are too full of "ad hoc" sources of inertia.

- These sources of inertia are there because without them, economic theory implies much faster reactions of economic variables to disturbances than we see in the data.

Motivation

Macroeconomics:

- Theorists complain that the Smets-Wouters model, and its descendants and ancestor (Christiano, Eichenbaum and Evans), are too full of "ad hoc" sources of inertia.

- These sources of inertia are there because without them, economic theory implies much faster reactions of economic variables to disturbances than we see in the data.

- Keynes just assumed prices move slowly. Neo-Keynesians "micro-found" Keynes by assuming there are costs of adjustment of prices, or "technological" barriers to frequent changes in them.

Motivation

Macroeconomics:

- Theorists complain that the Smets-Wouters model, and its descendants and ancestor (Christiano, Eichenbaum and Evans), are too full of "ad hoc" sources of inertia.

- These sources of inertia are there because without them, economic theory implies much faster reactions of economic variables to disturbances than we see in the data.

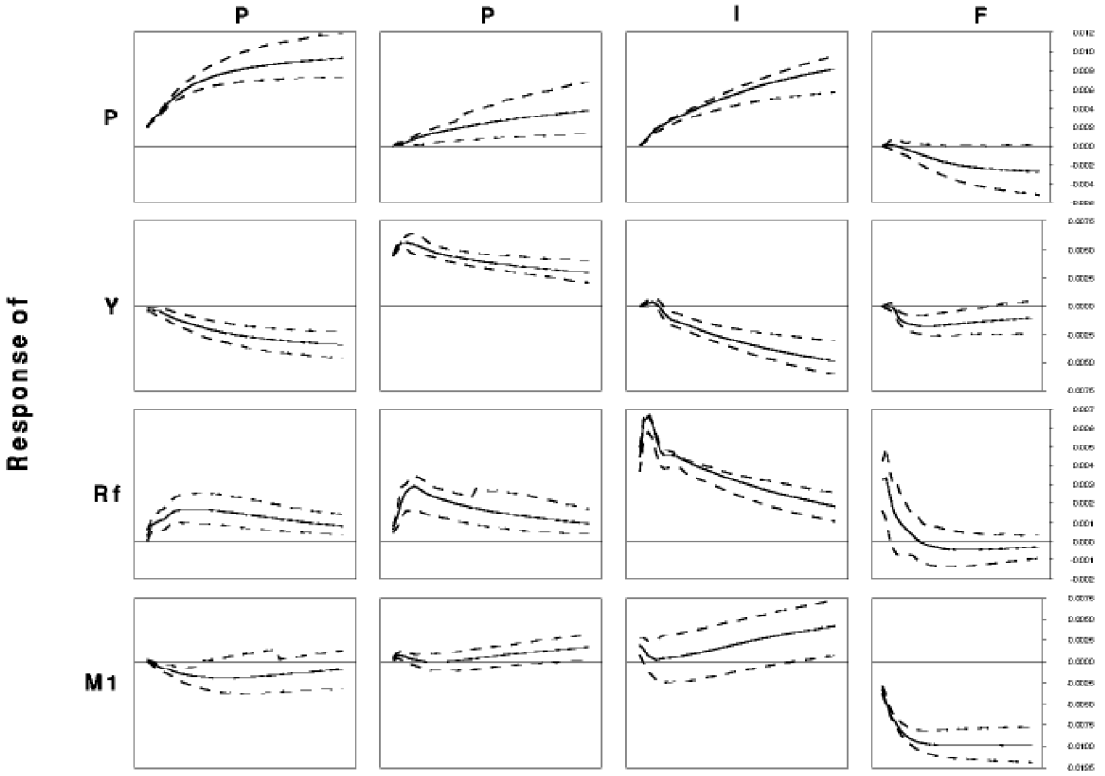
- Keynes just assumed prices move slowly. Neo-Keynesians "micro-found" Keynes by assuming there are costs of adjustment of prices, or "technological" barriers to frequent changes in them.

  But there is little empirical microfoundation for many of these costs.

Motivation

Figure 5

Four Variable Restricted Model



The plots are hard to explain with adjustment costs

- Responses on the diagonal show initial jumps, followed by smooth adjustment.
- The initial jumps imply that the variables *can* move abruptly, but the off-diagonal impulse responses, which reflect cross-variable effects, all show little or no initial effect, followed in some cases by sustained movement.
- "Adjustment costs" cannot easily explain this pattern.
- Rational inattention can.

The basic idea of rational inattention

- A person responding to randomly fluctuating market signals and translating them into actions has some of the characteristics of an engineer's communications channel: There are inputs (market signals), outputs (actions) and a maximum rate at which inputs can be translated into output without significant error.

- An internet connection has a **capacity** measured in bits per second (or, nowadays, more likely megabytes per second (MB/s or MiB/s)). Surely a person translating market signals to actions faces a similar, though probably tighter, limit.

Contrast with value of information in standard decision theory

Standard setup:

$$\max_{\delta} E[U(Y, X, Z, \delta)] \quad \text{subject to}$$

$$\text{either } \begin{cases} \delta \text{ a function of } X, Z \\ \delta \text{ a function of } X \end{cases} .$$

Difference in $E[U]$ between the two cases is the "value of the information in Z".

Entropy, discrete case

- **Entropy** is a measure of "how much uncertainty" there is in a probability distribution.

Entropy, discrete case

- **Entropy** is a measure of "how much uncertainty" there is in a probability distribution.
- It depends on the *distribution*, or probability measure, not on the values taken on by a random variable.
- If heads or tails each has probability .5, and next year's profits have probability .5 of being \$1 million or \$1.75 million, both random variables have distributions with the same entropy.
- It is $-\sum p_i \log(p_i) = -E[\log(p_i)]$, where i enumerates the points in the probability space.

Entropy, discrete case

Axiomatics: We assume first that we want the "amount of information" obtained by resolving the uncertainty in a distribution over a finite state space to depend only on the probabilities $\{p_i, i = 1, \ldots, n\}$ of the points in the space. Call this function $H(p_1, \ldots, p_n)$

Then we add the requirement that the uncertainty is additive in the following sense. Suppose $\{X_1, \ldots, X_m\}$ is a sequence of random variables defined on the n-point finite state space with probabilities defined by $\{p_1, \ldots, p_n\}$. Assume $Y = f(X_1, \ldots, X_m)$ is a function of these random variables with the property that the Y random variable has a unique value at each of the n points in the state space — meaning that observing Y, or the values of $\{X_1, \ldots, X_m\}$ is enough to reveal which point in the state space has been realized. After seeing X_1, our distribution over the state

space is $p(i \mid X_1)$. After seeing X_1 and X_2 it is $p(i \mid X_1, X_2)$, etc. Since observing Y is equivalent to observing the value of the state, we will use the notation $H(Y)$ as equivalent to $H(p_1, \ldots, p_n)$ and $H(Y \mid x)$ for the entropy of the distribution of Y conditional on a particular value x of the random variable X. $E_X[H(Y \mid x)]$, where the expectation is over the values of the random variable X, we write as $H(Y \mid X)$. We would like to have this property:

$$H(X_1) + H(X_2 \mid X_1) + \ldots H(X_n \mid X_1, \ldots, X_{n-1}) = H(Y)$$

This property is enough to tell us that $H(Y) = H(p_1, \ldots, p_n) = -E[\log(p_i)]$. We can choose the base of the log, but otherwise the measure is unique. If we use log base 2, the unit of measurement is called a "bit". One bit is the information in a flip of a fair coin. If we use natural logs the unit is called a "nat". Proving this result is not too hard, if one adds

the assumptions that H is continuous in $\vec{p}$ and that adding points to the probability space that have probability zero does not change the entropy.

Entropy, general case

- Entropy can be defined for distributions on any probability space.

Entropy, general case

- Entropy can be defined for distributions on any probability space.
- It is always defined relative to some base measure. For finite probability spaces, the base measure is counting measure and the density is the usual $\{p_i\}$.
- For continuously distributed random variables, The standard base measure is Lebesgue measure on $\mathbb{R}^n$, so entropy becomes $\int_{\mathbb{R}} p(x) \log(p(x))\, dx$.

Mutual information

- If X and Y are two random variables with a joint distribution defined by a density $p(x, y)$ over a base measure $\mu_x \times \mu_y$, then the **mutual information** between X and Y is $I(X, Y) = H(X) + H(Y) - H(X, Y)$.

Mutual information

- If X and Y are two random variables with a joint distribution defined by a density $p(x, y)$ over a base measure $\mu_x \times \mu_y$, then the **mutual information** between X and Y is $I(X, Y) = H(X) + H(Y) - H(X, Y)$.

- Equivalently, it is $H(Y) - H(Y \mid X)$ or $H(X) - H(X \mid Y)$. That is, it can be thought of as the expected reduction in the entropy of Y from observing X, or equivalently as the expected reduction in the entropy of X from observing Y.

- Note that $I(X, Y) = H(X)$ if and only if $H(X \mid Y) = 0$, i.e. if and only if there is no uncertainty about X after we observe Y.

Coding

Suppose we have a record of 1024 (2^{10}) time periods in which 5 events occurred. Our raw data is a string of 1024 0's and 1's, with the 1's in the positions corresponding to the dates of the events. We want to send this via a telegraph key that each quarter second can reliably transmit a 0 or a 1.

If we just send the raw data, it will take $1024/4 = 256$ seconds, or about 4 minutes. But of course we could instead agree with the person receiving the information that we are not sending the raw data, but just the positions in the sequence of the ones. Each such position can be represented as a number between 1 and 1024, and in binary notation these are 10-digit numbers. Since there are just five one's, we can get the message through with a transmission of 50 0's and 1's, or about 12 seconds.

We are **coding** the raw data into a different form that can be more efficiently transmitted through our telegraph key.

This coding scheme is a good one if the 1's are rare and randomly scattered through the sequence. If the 1's formed more than about 10% of the raw data, transmitting all their locations this way would take longer than transmitting the raw data. So our ability to code it to achieve faster transmission depends on the probability distribution from which the raw data is drawn.

If the message is always going to be five event dates, randomly drawn from the integers 1 to 1024, then the entropy of the distribution of messages is five times the entropy of a distribution that puts equal probability on 1024 points, i.e. $5 \log_2(1024 = 50$ bits. Or if the events are i.i.d. across dates, with probability 5/1024 of a one at each date, the entropy of each draw is .0445 and thus the entropy of the whole sequence is $1024 \times .0445 = 45.64$ bits.

Psychology experiments showing varying "capacities"

- Vertical vs. non-vertical, horizontal vs. non-horizontal, 45 degrees vs. not.
- 52+48 random dots , 52+48 dots nicely lined lup.
- RI assumes that if important information starts to come up as plots of (n,100-n) dots, the rational agent arranges that they come lined up, or even better, that a machine or assistant just announces "$n > 50$" or "$n < 50$".

The general static costly-information decision problem

$$\max_{f(\cdot,\cdot),\mu} E[U(X,Y)] - \theta I(X,Y) \text{subject to}$$

$$\int f(x,y)\, d\mu(x) = g(y), \quad f(x,y) \geq (0)\,.$$

- g is a given marginal density, relative to the base measure ν of the exogenous random variable y
- f is the joint density of the choice variable X and Y
- μ is the base measure for the distribution of X, in case it should turn out to be discrete.
- $I(X,Y)$ is the mutual information between X and Y.

More explicit mathematical structure

$$\max_{f(\cdot,\cdot),\mu} \int U(x,y) f(x,y)\, d\mu(x)\, d\nu(y)$$

$$-\theta \left(- \int \log\big(g(y)\big) g(y)\, d\nu(y) - \int \log\Big(\int f(x,y)\, d\nu(y) \Big) f(x,y)\, d\mu(x)\, d\nu(y) \right.$$

$$\left. + \int \log\big(f(x,y)\big) f(x,y)\, d\mu(x), d\nu(y) \le \kappa \right)$$

$$\text{subject to} \int f(x,y)\, d\mu(x) = g(y)\,, \quad f(x,y) \ge 0\,.$$

We could add other constraints. We've not been explicit about the domain of f.

FOC

$$\log\big(U(X,Y)\big)$$

$$= \lambda \left(1 + \log f(x,y) - 1 - \log \left(\int f(x,y')\, d\mu(y') \right) \right) + \psi(y)$$

or

$$q(y \mid x) = e^{\psi(y) U(x,y)/\lambda} .$$

This must hold at all points where $f(x,y) > 0$.

LQ Gaussian case

- If U is quadratic in x, y, the FOC's imply that we can choose a constant $\psi(y)$ that makes the form of $q(y \mid x)$ in the FOC Gaussian with variance $\lambda/2$.

- Since this problem has a concave objective function and convex constraints, a solution to the FOC's is a solution to the problem.

- What remains to check is the constraint

$$\int f(x, y)\, d\mu(x) = \int p(x) q(y \mid x)\, d\mu(x) = g(y)\ .$$

- With q the density of a normal with variance $\lambda/2$, this says that g is the density of a random variable X plus independent Gaussian noise with variance $\lambda/2$.

- Some g's satisfy this condition (e.g. a $N(0, \lambda)$) , but some can't (e.g. $U(0, 1)$, Gamma, $N(0, \lambda/4)$)

General static LQ problem

$$\max_{f(y,x)} E[-Y'AY + Y'BX - X'CX] - \theta\big(\log|\Sigma_Y| - \log\big|\Sigma_{Y|X}\big|\big)$$

subject to $\Sigma_Y - \Sigma_{Y|X}$ p.s.d. .

- We're using the result that the optimal conditional distribution of $Y \mid X$ is normal when the objective function is quadratic and Y is normal.
- Certainty equivalence applies: whatever information $\mathcal{I}$ we collect, once it's known, we can find the optimal X by replacing Y with $\hat{Y} = E[Y \mid \mathcal{I}]$ and solving the deterministic problem.

Rewrite objective function

Optimal $X = \frac{1}{2}C^{-1}B'\hat{Y}$, given current information, as can be verified easily from the FOC's. This lets us rewrite the objective function as

$$-\operatorname{trace}(\Sigma_{Y|X}A) - \operatorname{trace}(\Sigma_H A) + \frac{1}{4}\operatorname{trace}(\Sigma_H)BC^{-1}B' + \theta \log\left|\Sigma_{Y|X}\right| - \theta \log\left|\Sigma_Y\right|,$$

where we have used Σ_H for the variance matrix of $\hat{Y}$, which is also $\Sigma_Y - \Sigma_{Y|X}$.

Unconstrained solution

- If the "no-forgetting" constraint that $\Sigma_Y - \Sigma_{Y|X} = \Sigma_H$ be positive semi-definite does not bind, we can solve the problem analytically from the FOC's.

- The solution sets $\Sigma_{Y|X} = 4\theta(BC^{-1}B')^{-1}$.

Σ_H p.sd. constraint

- If the cost of information θ gets high enough, it becomes optimal to collect no information — i.e. choose $\Sigma_H = 0$. The distribution of X is then discrete, concentrated on a single point.

- In the univariate case, for this LQ problem, this is the only kind of discreteness possible.

- In the multivariate case it is also possible that θ is high enough to make collecting no information optimal, but there are intermediate possibilities, in which Σ_H is singular, and a simple analytic solution is unavailable.

Solution when p.s.d. contraint binds

- A numerical approach seems necessary.
- Write $\Sigma_H = VV'$, optimize numerically over V.
- If optimal Σ_H is singular, and V is taken to be square, the optimal V will be less than full rank and its individual elements are not uniquely determined.
- Some optimization programs will nonetheless find a correct solution for V efficiently, in which case you can just run the optimization program and check the rank of the resulting V.

- As a further check, you can then rerun the optimization with V $n \times m$, where n is the length of Y and m is the rank of V. If the rank is correct, this should make the optimization faster and more accurate.

Special cases: pure tracking, water-filling

- If the objective is simply $E[-(Y-X)'A(Y-X)]$, then it is easy to check that the optimal unconstrained solution is simply $\Sigma_{Y|X} = \theta A^{-1}$.
- If we further simplify, by specifying Σ_Y to be diagonal while $A = I$ in the pure tracking problem, we can see analytically how the rank of Σ_H changes with θ.
- The solution then moves from collecting no information, to collecting information only on the largest diagonal element of Σ_Y, then to collecting information on the largest two diagonal elements, keeping their posterior variances the same, then on the largest three, etc. as θ decreases, until finally we reach the unconstrained case where posterior variances on all elements of Y are the same ($\Sigma_{Y|X} = \theta I$).