RATIONAL INATTENTION AND MONETARY ECONOMICS

CHRISTOPHER A. SIMS

1. MOTIVATION

Everyone ignores or reacts sporadically and imperfectly to some information that they "see". I page through the business section of the New York Times most mornings, "seeing" charts and tables of a great deal of information about asset markets. I also most days look at `ft.com`'s charts of within-day movements of oil prices, stock indexes, and exchange rates once or twice. But most days I take no action at all based on this information I've viewed. In fact, if you asked me a half hour after I looked at the paper or the web site what the numbers were I'd viewed, I would usually be able to give at best a rough qualitative answer — unless there was some strikingly unusual data. If I were continually dynamically optimizing, I would be making fine adjustments in portfolio, spending plans, bill payment delays, etc. based on this information. It is intuitively obvious why I don't — the benefits of such continuous adjustment would be slight, and I have more important things to think about.

One might think that if we were to recognize that people don't use some freely available information, we would have to abandon optimizing-agent models of behavior. Some would be happy with this conclusion, but optimizing-agent models have served economic science well, so it is worthwhile asking whether it is possible to construct optimizing-agent models that are consistent with people not using

Date: January 17, 2015.

freely available information. "Rational inattention" models introduce the idea that people's abilities to translate external data into action are constrained by a finite Shannon "capacity" to process information. Such models do explain why some freely available information is not used, or imperfectly used.

Another appeal of such models is that they imply sluggish and erratic response of all types of behavior to external information. In macroeconomic data we see few examples of variables that respond promptly to changes in other variables. Keynesian models recognize inertia in prices, but in their simpler forms translate this inertia in prices into prompt and strong responses of quantities to policy and to other disturbances. This implication of Keynesian models can be softened or eliminated by the introduction of adjustment costs, but such costs are usually modeled one variable at a time and have little support in either intuition or formal theory. A rational inattention approach implies pervasive inertial and erratic behavior, and implies connections across variables in the degree and nature of the inertia.

Studies of transactions prices of individual products, which have proliferated in recent years as electronic cash registers have become common, show that prices tend to stay constant for extended periods of time, and to jump back and forth among a few specific price points when they do change. This pattern of discretely distributed prices is hard to reconcile with most existing theories of price sluggishness. Yet though this pattern was not part of the initial inspiration for rational inattention modeling, it has turned out that it is an implication of the rational inattention approach under fairly broad conditions.

In hopes that the reader is now interested in the topic, we turn to the basic mathematics of information theory.

2. INFORMATION THEORY

2.1. **Shannon's definition of mutual information.** Suppose we are sending the message "yes" and want to quantify how much information is contained in that message. Shannon's measure of information flow starts from the insight that the amount of information in that message depends on what *other* messages *might* have been sent instead. If the recipient of the message was already sure that the message was going to be "yes", no information at all is transmitted, and indeed no message need have been sent. If the recipient knew the message would be either "yes" or "no" and was unsure which, a small amount of information would be involved, and it would be easy to send it reliably. But if the recipient knew in advance only that the message would be some English language word, the message would contain much more information and would be much more difficult to send reliably. Shannon's idea was that the information transmitted ought to measured by how much the uncertainty of the recipient is reduced by receipt of the message.[1]

When two random objects, say X and Y have a joint distribution with a probability density function $p(x, y)$ Shannon's definition makes the mutual information between them

$$I(X,Y) = E[\log p(X,Y)] - E\left[\log\left(\int p(X,y)dy\right)\right] - E\left[\log\left(\int p(x,Y)dx\right)\right].$$

That is, the information between X and Y is the difference between the expected value of the log of the joint pdf of X and Y and the sum of the two expected values of the logs of the marginal pdf's of X and Y. This measure has some easily verified appealing properties. It is zero when X and Y are independent, and it is always

[1]Here we can only sketch the basic ideas of information theory. More complete treatments are in, e.g., Cover and Thomas (1991) or MacKay (2003).

non-negative. If we have a sequence of observations, say on Y and on Z, we would like the information about X in seeing Z, then Y, to be the same as that in seeing Y, then Z. Thus we would like $I(X, Y)$, calculated from the joint distribution for X and Y, plus $I(X, Z \mid Y)$, calculated using the joint pdf of X and Z conditional on Y, to be the same as $I(X, Z)$ plus $I(X, Y \mid Z)$. It turns out that these simple properties are restrictive enough to leave us with essentially *only* the Shannon measure of mutual information. The "essentially" is needed because we have not specified the base of the log function in the definition. The usual base is 2, in which case the unit of information is a "bit", while sometimes it is convenient to use base e, in which case the unit is called a "nat". [2]

Besides these intuitively appealing properties, the Shannon measure stands out for its proven usefulness in communications engineering. These days, most people are familiar with the idea that they can have fast or slow internet connections, that there is a measure for the speed (megabits or megabytes (1 byte = 8 bits)) per second), and that the measure doesn't depend on either the content of the messages being sent (music, text, pictures) or on the physical details of the connection (fiber optic, cable, DSL, etc.).

We should note that the symmetric definition given above is equivalent to

$$I(X, Y) = E\left[E[\log(q(X \mid Y)]\right] - E[\log(h(X))]\,,$$

where $h(X) = \int p(X, y)\, dy$ is the marginal pdf of X and

$$q(X \mid Y) = p(X, Y) / \int (p(x, Y) dx$$

[2] See Bierbrauer (2005, chapter 8) for further discussion of the uniqueness.

is the conditional pdf of $X \mid Y$. The quantity $-E[\log(h(X))]$ is called the **entropy** of the random variable X, so that this form of the definition of $I(X,Y)$ makes it the expected reduction in entropy of X from seeing Y. The symmetry of the first definition makes it clear that the expected reduction in entropy of Y from seeing X is the same as the expected reduction in the entropy of X from seeing Y.

2.2. **Channels, capacity.** Shannon defined a channel as a description of possible inputs and of conditional distributions of inputs given outputs. For example, an ideal telegraph line could send a "dot" or a "dash" (the inputs) and produce a dot at the other end when the input was a dot, and a dash when the input was a dash. A more interesting channel would be a noisy telegraph line, in which the a dot or dash input reproduces itself in the output only with probability .6, otherwise producing the opposite. In this latter channel, in other words, the probability of error is .4 with each transmission. Or a channel might be able to send arbitrary real numbers x drawn from a distribution with variance no greater than 1, producing in the output $y \sim N(x, \sigma^2)$.

The channel only defines conditional distributions of outputs Y given inputs X. The mutual information between inputs and outputs depends also on the distribution of the inputs. If we choose the distribution of the inputs to maximize the mutual information between inputs and outputs, the channel transmits information at its **capacity**. The ideal telegraph key makes the distribution of inputs given outputs degenerate, with all probability on the true value of the input. A discrete distribution with probability one on a single point has entropy 0 ($0 \cdot \log(0) + 1 \cdot \log(1)$, with the convention that $0 \cdot \log(0) = 0$, the limiting value of $a \cdot \log(a)$ as $a \downarrow 0$). The information flow is maximized if the input makes dots and dashes equally

probable, in which case it is one bit per time period. The noisy telegraph key also has maximal mutual information between input and output when the dashes and dots are equiprobable in the input. Then the information flow rate is .029 bits per time period. The channel with Gaussian noise has maximal information flow rate when the input is distributed as $N(0,1)$, in which case the information flow rate is $-\frac{1}{2}\log_2\left((\sigma^2/(1+\sigma^2)\right)$ bits per time period. When the noise is as variable as the input, so $\sigma^2 = 1$, for example, the rate is .5 bits per time period.

2.3. **Coding.** It is a relatively familiar idea these days that one can take information in various forms and transmit it via an internet connection. Many of these connections naturally take "ones" and "zeros" (commonly called bits, though this is not exactly the same as the information theory use of that term) as input, and computer disk files represent any kind of information as a pattern of bits. The well-known ascii code maps each number or upper or lower case letter into a pattern of seven bits. Pictures can be mapped into bit patterns that describe pixels — color intensity amounts at specific points in the picture. This kind of translation of diverse types of information into bits is coding.

But there are many possible ways to map letters and numbers or picture descriptions into bits. Text translated into ascii codes generally does not emerge with serially uncorrelated bit patterns or with equal numbers of 0's and 1's, and as a result is not ideal input for our ideal telegraph key. There are algorithms that translate such inefficiently coded files into more efficiently coded ones — for example the zip (for general files) and jpeg (for image files) compression schemes that most computer users have encountered. These compression algorithms produce patterns of zeros and ones that are more nearly i.i.d. and mean .5, and thereby become smaller files.

The shrinking of these files is equivalent to making them transmit more quickly through an ideal telegraph key.

The coding theorem of information theory states that regardless of the nature of the input we wish to transmit, it can be "coded" so that it is sent with arbitrarily low error rate at arbitrarily close the the channel capacity transmission rate. To get an idea of what coding is and of the meaning of the theorem, suppose we are sending a simple bit-mapped graph of a few black and white lines. The graph has been scanned into a 100×100 grid of pixels, and the file we wish to send is the 100 rows of pixels, one row at a time. With a 0 representing white and a 1 representing black, most of the file will be zeros. Our channel is a perfect telegraph key. Say two per cent of the file is 1's. If we simply send the raw file through the channel, it will take 10,000 time periods, one for each pixel. But we could instead transform the file so that a 0 now represents the sequence 000, while 1001 represents 001, 1010 represents 010, etc. (Note we end up not using 1000 at all.) Then $.98^3 = .94$ of our three-pixel blocks will be represented by a single 0 in the output, while .06 of them will be represented by four-element sequences. On average, our three-pixel blocks will take $.94 \times 1 + .06 \times 4 = 1.18$ time periods to transmit, so the whole file will take $10000 \times 1.18/3 = 3934$ time periods to transmit. If we think of the file as drawn from a collection of files that have i.i.d. sequences of zeros and ones with probability .02 of a one, the entropy of the file is $10000(.02 \log_2(.02) + .98 \log_2(.98) = 1414$ bits.[3] If we use the proposed coding, then, we would be sending $1414/3934 = .36$ bits per time period, whereas as we have already noted the channel capacity is 1 bit

[3]If we were really considering only graphics files with black and white line art, the zeros and ones would not actually be i.i.d. (because the ones occur in mostly continuous lines), so the entropy would be smaller and faster transmission possible.

per time period. To get closer to the channel capacity would require more elaborate codes, for example using blocks longer than three.[4].

This example may also help in understanding an important and possibly confusing fact: Even though our ideal telegraph line transmits without error and at a finite rate, a channel that takes continuously distributed input cannot transmit without error unless it has infinite capacity. Suppose input X can be any real number, and output Y simply equals X. Consider our 10000-pixel graphic file above. If we take its sequence of zeros and ones and put a decimal point in front of it, it becomes the binary representation of a real number between zero and one. We could then send it through our channel in a single time period without error, a rate of 1414 bits per time period. And of course the same idea would work no matter how large the file, so there is no upper bound on the transmission rate.

The coding theorem is not constructive. Given a channel and a type of message to be sent, finding a way to code it so it can be sent at close to capacity is generally difficult and has generated a substantial literature in engineering.

Our example of coding above illustrates another complication that we will be mostly ignoring in what follows: coding introduces delay. We showed how to send a file that is mostly zeros by sending the message in blocks. But to do this we need to wait until we have a full block to transmit, which generates some delay. How much delay depends on the nature of the channel and of the message — that is, on properties of the channel and message beyond the channel capacity and the entropy of the message. We ignore coding delay for two reasons: we are at this stage in applications to economic behavior trying to avoid needing to discuss the physical

[4]A longer-block coding example is in the appendix to my 1998 paper.

characteristics of people as information channels; and coding delay is likely to be small — the proportional gap between channel capacity and actual transmission rate decreases at least at the rate $1/n$, where n is the block length of the coding (Cover and Thomas, 1991, section 5.4).

3. Information theory and economic behavior

The idea of rational inattention is to introduce into the theory of optimizing agents an assumption that their translation of observed external random signals into actions must represent a finite rate of information flow; that is, economic agents are finite-capacity channels.

Before we proceed to discussing rational inattention models, we should note that these models do not subsume or claim to replace all previous economic models of costly information. In statistical decision theory it is possible to quantify the utility value of observing a random variable, and if the problem includes a budget constraint, to convert this value into a dollar equivalent. This kind of "value of information" applies when there is some physical cost to acquiring the observation — commissioning a marketing survey, drilling a test well, etc. This kind of information cost has nothing to do with the number of bits of information acquired by observing the random variable. Finding whether a test well indicates oil is present may cost thousands of dollars, yet provide only the answer to a yes-or-no question — i.e. no more than one bit of information. Rational inattention theory provides no guidance on whether drilling a test well is a good idea. Where it might provide guidance is in explaining why an executive in the oil company, having had a report on the test well on her desk along with other reports about routine matters, might after "looking at" all the reports seem to know the test well report in detail, while

having only a vague idea of what was in the other reports. The test well report was important to her job, the others less so, so the others are absorbed less precisely.

Notice also that in the examples that follow the information flow rate is lower than any reasonable guess as to human beings' actual Shannon capacities. It is probably most natural to think of an abstract economic agent as having a shadow value of capacity rather than a fixed capacity bound, because economic optimizations in fact represent only a tiny part of the information-processing that people do. To get realistic delay and noisiness in reactions to information in models where economic decision-making is the only reason to process information, we need to postulate very low Shannon capacity, yet at small costs of capacity we find optimizing agents use little of it. This reflects the well-known fact brought out by Akerlof and Yellen (1985) that in the neighborhood of an optimum, modest deviations from fully optimal choices are likely to have very small consequences. People may use economic information at a low rate not because they could not possibly use it more precisely, but because the benefits of doing so would be small and there are other important uses of information-processing capacity.

3.1. **The Gaussian case.** Rational inattention models are easiest to handle when random variables are all jointly normal. The entropy of a k-dimensional $N(\mu, \Sigma)$ random vector is $\frac{1}{2}(\log(2\pi) + \log|\Sigma| + k)$. This means that the mutual information between two jointly normally distributed random vectors X and Y is half the difference between the log of the unconditional covariance matrix of Y and the log of the residual covariance matrix for a regression of Y on X. It depends only on the correlation matrix of X and Y, not on the levels of the variances themselves. If X and Y are each one-dimensional, their mutual information is just $-\frac{1}{2}\log(1-\rho^2)$,

where ρ is the correlation of X with Y:

$$X, Y \sim N(\mu, \Sigma) \Rightarrow I(X,Y) = \tfrac{1}{2}(-\log|\Sigma| + \log(\mathrm{Var}(X)) + \log(\mathrm{Var}(Y)) = -\tfrac{1}{2}\log(1-\rho_{XY}^2)\,.$$

Joint normality of a signal Y and an action X is a strong assumption, because rational inattention theory naturally takes the distribution of Y as given and then, based on the loss function and the information constraint, implies a joint distribution for X and Y. Generally, even with Y normally distributed, the information-constrained optimal joint distribution for Y and X is not normal. A comforting result is that there is a form for the loss function that implies joint normality as the optimal form of the joint distribution.

A general static information-constrained decision problem can be formulated as follows:

$$\max_{f()}\{E[U(X,Y)] = \int U(x,y)f(x,y)\,dx\,dy \qquad \text{subject to}$$

$$(\dagger) \qquad \int f(x,y)dx = g(y) \quad \text{all } y$$

$$f(x,y) \geq 0 \quad \text{all } x,y$$

$$I(X,Y) = \int \log(f(x,y))f(x,y)\,dx\,dy$$

$$- \int \log(\int f(x,y')dy')f(x,y)dy\,dx - \int \log(g(y))g(y)dy \leq \kappa\,,$$

where X is the choice variable, g is the given marginal pdf of Y and κ is the maximum information flow rate between Y and X. The objective function is linear in the object of choice (f) and the constraint set is convex, so the problem has a unique maximal value for the objective function. A closely related formulation (actually applied in the examples we will take up) assumes that capacity is variable, at a

cost. The left-hand side of the information constraint then appears in the objective function, multiplied by the cost, rather than in a separate constraint.

It may be puzzling that the agent is modeled as choosing a joint distribution rather than as simply choosing X. The problem could be formulated equivalently by saying that the agent chooses an observation $Z = h(Y, \zeta)$, where ζ is a random variable independent of Y and h is an arbitrary (measurable) function. The information constraint is $I(Z, Y) \leq \kappa$ and the agent chooses also a function $d()$ and sets $X = d(Z)$. Here the choice of information and the setting of X are separated, which may perhaps be easier to understand. But this formulation is equivalent to the one in terms of choosing f, and has the disadvantage that the same solution $f()$ can generally be characterized with many different $d(), h()$ pairs.

At points in X, Y-space where $f(x, y) > 0$, the first-order conditions for an optimum require

$$U(x, y) = \lambda \left(\log(f(x, y)) - \log(\int f(x, y) dy) \right) - \mu(y) \,, \tag{1}$$

where λ is the Lagrange multiplier on the information constraint and $\mu(y)$ is the Lagrange multiplier on the constraint that defines the marginal distribution of Y. This condition can be rearranged to read

$$p(y \mid x) = M(y) e^{\frac{1}{\lambda} U(x,y)} \,. \tag{2}$$

Since conditional density functions integrate to one, (2) in turn implies

$$\int M(y) e^{\frac{1}{\lambda} U(x,y)} \, dy = 1 \,, \text{ all } x \,. \tag{$*$}$$

Suppose $U(x, y)$ is quadratic in x and y jointly and y is itself distributed as $N(0, \sigma_y^2)$. To keep the algebra simpler, we will assume it is homogenous, i.e. of

the form

$$U(x,y) = -(\omega_{xx}x^2 + 2\omega_{xy}xy + \omega_{yy}y^2)\,.$$

To keep the problem non-trivial, we require ω_{xx} positive. (Otherwise the solution is to make x arbitrarily large.) We can, for any given λ, satisfy equation $(*)$ by choosing

$$M(y) = C\exp\left(\left(\omega_{yy} - \frac{\omega_{xy}^2}{\omega_{xx}}\right)y^2\right)\,,$$

where C is some positive constant. This makes the integrand in $(*)$, as a function of y, proportional to the pdf of a

$$N\left(\frac{\omega_{xx}}{\omega_{12}}x, \frac{\lambda\omega_{xx}}{\omega_{xy}^2}\right)$$

Distribution. Since the variance of this distribution does not depend on x, the integral of the pdf is the same for all x, and we can therefore choose C to satisfy $(*)$ for all x. It remains to check the constraint that the marginal distribution of y match the given distribution. Since the integrand in $(*)$ defines the conditional density of $Y \mid X$, we must have

$$(\dagger) \qquad \frac{\lambda\omega_{xx}}{\omega_{xy}^2} \leq \sigma_y^2\,.$$

That is, the solution cannot imply that we know less about y after collecting information than we did originally. In a version of this problem where λ is an exogenously given cost of information, this requirement implies that for a large enough cost of information, it will be optimal not to choose the distribution of x to have full support (in fact it is then optimal to collect no information, making x constant.) But with a capacity constraint, this condition tells us only that no matter how small the capacity available, the Lagrange multiplier on its constraint remains bounded above.

We can make the joint distribution of X and Y normal by giving X a normal distribution and multiplying its pdf by the conditional pdf for $Y \mid X$ that we have constructed. The mutual information between Y and X for a joint normal distribution is $-\frac{1}{2}\log(\text{Var}(Y)/\text{Var}(Y \mid X))$. Therefore the mutual information constraint uniquely determines λ, since the conditional variance of Y is proportional to it. For any value of λ implying strict inequality in (†), we can make the unconditional variance of y implied by our constructed joint distribution match the given σ_y^2 by appropriate choice of σ_x^2. To be specific, we choose

$$\sigma_x^2 = \frac{\omega_{xy}^2}{\omega_{xx}^2}\sigma_y^2 = \frac{\lambda}{\omega_{xx}},$$

which is always positive when (†) is satisfied.

So we have shown that all the first-order conditions for an optimum can be statisfied by a joint normal distribution for Y and X. Since the objective function is linear in f and the constraints are all convex in f, we can be sure that the joint normal distribution is a solution.

3.2. **Some qualitative conclusions, based on Gaussian-Linear-Quadratic examples.** The appendix describes how to solve general linear-quadratic optimal control problems. Here we apply the method laid out there to some simple examples that provide insight into the economic implications of rational inattention.

3.2.1. *Rational inattention smooths responses and injects signal-processing noise.* Suppose P_t is an asset price and X_t is some action an agent takes in response to the asset price. Suppose that in the absence of an information constraint the optimal way to set X_t is to set $X_t = P_t$. If P is a Gaussian stochastic process then, unless it is

constant, $P_{t+s} \mid \{P_s, s < t\}$, the distribution of P_{t+s} given the history of P up to time t, is a Gaussian random variable.

If the optimal choice of X without an information constraint would be $X_t = \alpha P_t$, it is impossible to implement this choice under RI, because it makes knowledge of X_{t+s} completely resolve the continuously distributed uncertainty about P_{t+s}, which as we have already observed implies an infinite information flow rate. And it is not enough simply to add noise. Suppose $X_t = \theta P_t + \varepsilon_t$. Continuously traded asset prices tend to behave like Wiener processes over small time intervals. In particular, the variance of $P_{t+\delta} - P_t$ decreases linearly with δ and price changes over non-overlapping time intervals are independent. If ε_t also has this character, then the correlation of $X_{t+\delta} - X_t$ with $P_{t+\delta} - P_t$ tends to some non-zero level as δ shrinks. But that means that the mutual information between $P_{t+\delta} - P_t$ and $X_{t+\delta} - X_t$ tends to a constant as δ shrinks. Thus given a fixed time interval we can, by slicing it up into arbitrarily small subintervals, convey arbitrary amounts of information in the fixed time interval.

It is common to represent continuous time Gaussian processes as stochastic differential equations, of the form

$$dy_t = g(y)dt + h(y)dW_t\,, \tag{3}$$

where dW_t is a vector of independent white noise processes. The kind of argument we have given above implies that if y consists of two components, $y = (x, z)$, and if $h(y)$ is full rank, then for the rate of information flow between z and x to be finite, $h(y)$ must be block diagonal, with blocks corresponding to x and z. This implies that over short time intervals x and z each have variation dominated by their own disturbance process. The component of, say, x that is related to the z shock process

must be "more differentiable" than the component related to x's own shock process, so that the variance of changes in x can be dominated by the own-shocks at small time time intervals.

3.2.2. *RI solutions are a special case of rational expectations with noisy observations.* Consider this very simple dynamic tracking problem. We have a target process y_t that is a first-order univariate autoregression, and we wish to keep our action x_t close to it, with quadratic losses. We can tighten our variance for y_t before we choose x_t by paying an information cost of λ per nat. Formally,

$$\min_{x_t,\sigma_t} \tfrac{1}{2} E\left[\sum_{t=0}^{\infty} \beta^t \left((y_t - x_t)^2 + \lambda \log\left(\frac{\rho^2\sigma_{t-1}^2 + \nu^2}{\sigma^2}\right)\right)\right] \quad \text{subject to} \tag{4}$$

$$y_t = \rho y_{t-1} + \varepsilon_t\,, \tag{5}$$

where $\nu^2 = \mathrm{Var}(\varepsilon_t)$, σ_t^2 is the variance, after information collection, at t for y_t, and therefore $\rho^2\sigma_{t-1}^2 + \nu^2$ is the variance for y_t based on time $t-1$ information, before collecting information at time t.

It is clear that it will be optimal to make x_t always the expectation of y_t given information at t, so we can reduce the problem to one in which the only choice variable is σ_t^2:

$$\max_{\sigma_t} \sum_{t=0}^{\infty} \beta^t \left(\sigma_t^2 + \lambda \log\left(\frac{\rho^2\sigma_{t-1}^2 + \nu^2}{\sigma^2}\right)\right)\,. \tag{6}$$

This problem can be solved by standard methods, and it has a solution in which σ_t^2 is constant at some finite value. As one might expect, $\sigma_t^2 \to 0$ as $\lambda \to 0$. Also, $\sigma_t^2 \to \infty$ as $\lambda \to \infty$. This latter result brings out the fact that we have ignored to this point the requirement that $\sigma_t^2 \leq \rho^2\sigma_{t-1}^2 + \nu^2$. That is, one cannot improve the objective function by "forgetting" previously known information about y. So the

full solution is that if the solution to the unconstrained problem implies violation of this forgetting constraint, no information is collected and uncertainty about y is allowed to grow. If the variance of uncertainty about y grows to the point where it exceeds the variance of y in the unconstrained solution, the "no-forgetting" constraint ceases to bind and the solution path begins to follow the unconstrained solution.

Considered as a univariate process, x_t inherits the properties of y_t. This is a general characteristic of RI (and other noisy-observation rational expectations) dynamic optimizations: relative to the decision-relevant information set, the decision variables have the same dynamic structure as would the decision variables in the problem with no information-processing constraint. (Here the no-constraint solution would just be $x_t = y_t$.) It is easy to see that, denoting information available to the decision-maker at time t by $\mathcal{I}_t$, $E[x_t \mid \mathcal{I}_{t-1}] = E[E[y_t \mid \mathcal{I}_t] \mid \mathcal{I}_{t-1}] = \rho x_{t-1}$, so that x_t is an AR process with the same parameter as y. But even though the best predictor of x_t from its own past is ρx_{t-1}, this is generally not the best predictor of x_t from the joint past of y and x.

What then is the joint times series behavior of x_t and y_t in the unconstrained solution? The prediction error for y_t based on information available to the decision maker at time $t-1$ is $y_t - \rho x_{t-1}$. The choice of x_t will be based on an improved estimate of this error, and since everything is jointly Gaussian. we can write

$$x_t = \rho x_{t-1} + \theta(y_t - \rho x_{t-1}) + \xi_t \,, \tag{7}$$

where ξ_t is pure time-t information-processing error and therefore uncorrelated with $\{y_{t-s}, s \geq 0\}$ or with $\{x_{t-s}, s \geq 1\}$. This lets us derive a joint autoregressive

representation of (y, x) as

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \rho & 0 \\ \theta\rho & (1-\theta)\rho \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \theta\varepsilon_t + \xi_t \end{bmatrix} . \tag{8}$$

This implies the moving average representation

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \sum_{s=0}^{\infty} \begin{bmatrix} \rho^s & 0 \\ \theta\rho^s \sum_{u=0}^{s}(1-\theta)^u & \rho^s(1-\theta)^s \end{bmatrix} \begin{bmatrix} \varepsilon_t \\ \theta\varepsilon_t + \xi_t \end{bmatrix} . \tag{9}$$

Notice that if the time unit were very small, we would expect ρ to be near one and, to be consistent with small information flow over small time intervals, θ to be near zero. Then the second diagonal component of the sequence of weighting matrices in (9) is the weights on the noise component, and the lower left off-diagonal component is the weights on the part of x that is related to y. We see that as our reasoning above implied, the systematic part of x has small weight ($\theta\rho$) on the initial shock, but that the weight rises smoothly, nearly linearly at first, as we go to more distant lags of the shock. The noise component responds immediately, and the weights decline rapidly — it is less serially correlated than y itself, while the systematic part of x is much more serially correlated than y itself.

Note also that this solution is exactly what we would have obtained if we simply postulated that the optimization has to be based on observing at each t a noisy indicator variable $z_t = y_t + \zeta_t$, The variance of ζ would determine the corresponding value of θ in the expressions above, and $\theta\zeta_t = \xi_t$. What is added by the derivation from rational inattention is i) that the RI theory predicts that θ and the variance of ξ_t will vary systematically if ν^2 (the variance of ε_t) or λ changes and ii) we can show that the normal distribution for the "measurement error" is actually what an

agent will optimally choose with this objective function. If we made y multivariate we would have still further implied restrictions on the relation of information processing error to underlying disturbance processes and to the objective function.

We were able to solve this problem in two steps. First we recognized that, regardless of the error variance, it was going to be optimal to set $x_t = E[y_t \mid \mathcal{I}_t]$. That allowed us to convert the problem into one that involved only choice of error variance matrices. This two-step process is possible generically in LQ RI problems: First solve for the optimal function relating control variables to states, using certainty equivalence. Then use that solution to find the objective function value as a function of the sequence of error variance matrices alone. The first stage is a standard LQ control problem. The second stage is nonlinear, but deterministic.

Finally, observe that we had to take account of the $\sigma_t^2 \leq \rho^2\sigma_{t-1}^2 + \nu^2$ constraint, and this slightly complicated our solution. In a multivariate problem the corresponding constraint is that the time-$(t-1)$ covariance matrix for the state at t minus the post-observation covariance matrix must be a positive semi-definite matrix. Imposing this constraint, when it is necessary to do so, can be much more complicated than imposing it in a univariate problem.

3.2.3. *Rational inattention creates correlation across initially independent sources of uncertainty.* In our LQ dynamic tracking problem that reduces to (6), suppose there is no serial correlation, i.e. $\rho = 0$. Then the solution is obviously just $\lambda = \sigma^2$. But now add the complication that in fact $y_t = \sum_i z_{it}$, where $z_{it} \sim N(0, \omega^2)$, independent across t and i. Brief reflection makes it clear that this complication is no complication at all. For optimally choosing x in the face of information process costs, all that matters is that $y_t \sim N(0, n\omega^2)$, where n is the number of elements in z. Note,

though, that this implies that even if the vector z is freely observable, it will be optimal to collect information only about $\sum_i z_{it}$. The variance of any linear combination $c'z_t$ of the z_{it}'s that is uncorrelated with $\mathbf{1}'z_t$ will not be reduced, no matter how low the information cost parameter λ. This implies that the conditional distribution of z_t after an observation has been taken will be of the form

$$\omega^2(I - \alpha(1/n)\underset{n\times n}{\mathbf{1}}),$$

where $\alpha = 1$ when $\lambda = 0$ and $\alpha \to 0$ as $\lambda \uparrow n\omega^2$. Even though the uncertainty about z_t was uncorrelated across elements of the z_t vector to start with, it optimally becomes negatively correlated across i after information processing.

While this point may seem obvious, taking account of it can complicate analysis. It can be attractive for analytic convenience to assume that uncertainty is constrained to be reduced so as to keep the correlation structure[5] of the z's the same before and after observations are taken. This amounts to discarding one of the important insights from rational inattention theory, however, and should be seen as a last resort at best.

3.2.4. *Rationally inattentive agents react more slowly to slowly-moving components of an aggregate.* A very stylized model of pricing behavior might have a monopolist trying to match prices to a linear function of costs, with quadratic losses. Suppose cost is the sum of two components, one fast-moving, a univariate autoregression with lag coefficient (for example) .4, and another slow moving component with lag coefficient .95. Suppose we make the innovation variances to these two components independent of each other and pick them so that the unconditional variances of the

[5]More precisely, the eigenvectors of z's covariance matrix.

two components are equal. We also assume future costs are discounted at the rate β. Formally, the problem is

$$\min_{p,\Sigma} E\left[\sum_{t=0}^{\infty}\beta^t(\mathbf{1}'\Sigma_t\mathbf{1}+\lambda(\log(|\Omega_{t-1}|)-\log(|\Sigma_t|)))\right] \quad \text{subject to} \tag{10}$$

$$\Omega_t = \rho\Sigma_t\rho' + \nu \tag{11}$$

$$\Omega_t - \Sigma_t \text{ positive semi-definite}, \tag{12}$$

where our example numbers make

$$\rho = \begin{bmatrix} .95 & 0 \\ 0 & .4 \end{bmatrix}, \quad \nu = \begin{bmatrix} .0975 & 0 \\ 0 & .86 \end{bmatrix} \tag{13}$$

and λ is the cost of information. As might be intuitively clear, since the maximizer cares only about the sum of the two components, when information costs are low he will choose to make the variances of the components conditional on his information roughly equal and negatively correlated. Since the innovation variance for the slow-moving component is smaller, it is optimal not to track the innovation variance of that component closely, but rather to allow uncertainty about that component to cumulate until it approaches that in the fast-moving component. With $\beta = .9$ and $\lambda = 1$, our example makes the optimal choice

$$\Sigma_t = \begin{bmatrix} 0.373 & -0.174 \\ -0.174 & 0.774 \end{bmatrix}, \tag{14}$$

from which we see that the post-observation variance of the fast-moving component is 8% smaller than its innovation variance, while that of the slow-moving component is nearly four times larger than its innovation variance. When we relax the

information constraint by setting $\lambda = .1$, we find instead

$$\Sigma_t = \begin{bmatrix} 0.318 & -0.300 \\ -0.300 & 0.380 \end{bmatrix}. \tag{15}$$

"News" about the fast-moving component is perceived fairly promptly, while there is little immediate reaction to news about the slow-moving component. The uncertainty about the two components is ex post negatively correlated, reflecting the fact that the monopolist cares only about the sum of the two components and chooses to have imprecise knowledge about how the sum is allocated across components. And as the information constraint is relaxed, it is applied more to the fast-moving than to the slow-moving component.

3.2.5. *Losses from imperfect information processing are small, implying that even small information costs are likely imply substantial imprecision in reactions to signals.* In these examples, information-processing noise increases linearly with variance. The standard deviation of information processing noise therefore increases very rapidly with information processing costs in the neighborhood of zero processing costs. Though our examples have not been realistically calibrated, when models are realistically calibrated (e.g. Luo (2008)) small information costs lead to low optimal information flow rates and substantial effects on dynamic behavior.

3.3. **Contrast with Mankiw-Reis formulation.** In an influential paper Mankiw and Reis (2002) proposed a way to model inertial behavior that they call "sticky information". They discuss their approach in their contribution to this Handbook (Mankiw and Reis, 2010). Their work is motivated by some of the same insights that motivate the rational inattention approach. They postulate that agents update their information only at regular intervals that are either fixed or (in later work)

variable at a cost. At an information update, agents formulate plans for the period until the next update and stick with those plans. This implies delay and imprecision in response to variation in market signals, just as does rational inattention.

Their formulation is somewhat easier to incorporate into standard macro models, but it is quite different in many of its implications from rational inattention, and it takes us less far along the road away from ad hockery. At updates, agents see all the random variables that define the state of the economy, which are generally taken to be continuously distributed, without error, which as we have seen implies an infinite information flow rate. In a rational inattention setting, no continuously distributed external source of random variation is ever perceived without error, even with a lag. Under RI, delays in reacting to information depend on the amount of serial correlation and the size of disturbances to the external variable; when the external variable moves slowly and varies little, delays in reacting to it can be very long. Under sticky information, there is no such connection of the nature of the external variation to the amount of delay in reacting to it.

RI, as we have seen, has rich implications about how information from multiple sources is perceived and about how the relative precision of information about different variables depends on loss functions and on the stochastic structure of the external variation being tracked. Sticky information implies no theory about relative precision or delay in observation of different variables. It can allow for differences across variables by allowing for the rate of information collection to be different for different variables, but such formulations are less tractable.

Sticky information implies a different approach to possible microeconomic empirical verification of the theory. It suggests that we would want to examine how

often firms or individuals change "plans" for behavior and use these frequencies as an index of the effects of information constraints. Rational inattention, on the other hand, implies that behavior may continually but imprecisely be reacting to external signals, even when information effects are strong. As we will see below, outside the linear-quadratic Gaussian framework rational inattention can imply behavior that changes only at discrete intervals, yet at the same time imply that imprecise knowledge of the state prevails as much at change dates as at other dates.

3.4. **Beyond LQ.** Sims (2006), Matějka (2009), Matějka (2008), and Matějka and Sims (2009) take up models in which objective functions are not necessarily quadratic and supports of distributions are not necessarily unbounded. This necessarily takes us out of the realm of certainty equivalence and Gaussian distributions. Probably the most interesting result emerging from this work is that solutions often imply a discrete distribution for agent actions, even when the external uncertainty is continuously distributed. The result is the outcome of numerical calculations in most of these papers, but Matějka and Sims (2009) provides an analytic result covering a fairly broad category of models. They show that if i) the objective is to maximize $U(|x-y|)$, with U having its maximum at zero, ii) U is analytic on the entire real line, and iii) the given marginal distribution of y has bounded support, then with any cost on mutual information between x and y, the marginal distribution of x is optimally concentrated on a finite set of points.

This kind of result is interesting, because micro-economic data on product prices show not only that prices stay constant over moderately long time intervals, but also that when they change they often jump among a finite set of values (Eichenbaum, Jaimovich, and Rebelo, 2008). There are a number of models in the literature

that can explain why prices might change only occasionally, but none that explain why, when they do change, they should move among a discrete set of values. Rational inattention provides an explanation.

If rational inattention is playing even a partial role in determining price-setting behavior, it casts into doubt interpretations often placed on price micro-data. Rationally inattentive price setters do not fully adapt to all available information each time they change their prices. Intervals between price changes are therefore nearly irrelevant in determining the degree to which pricing responses to external information (e.g. monetary policy) are delayed or incomplete.

3.5. **General equilibrium.** Up to this point we have been discussing models of the behavior of individuals reacting to "external" information sources. In modeling an entire economy, or even a market, we must consider interacting agents. This raises special difficulties, as standard market equilibrium models assume prices adjust to clear markets. In a model of a competitive market, prices are usually taken as "external" to both suppliers and purchasers, and it is assumed that both sides of the market see and react to the price. That is how markets are assumed to clear. But in reality prices vary stochastically. If both sides of the market react to market prices with rational inattention, then neither side is reacting precisely and immediately. Prices therefore cannot play their usual market-clearing role.

There are a few models in the literature that consider markets with rationally inattentive agents. They do so by allocating variables to agents, with each variable a decision variable for one type of agent and an external signal to others. For example, Matějka (2009) considers a market with a monopolistic seller choosing prices subject to an information constraint on tracking costs. In a companion paper 2008

he considers a monopolistic price setter facing consumers who face an information constraint on tracking prices. Maćkowiak and Wiederholt (2009a) have set out a complete dynamic stochastic general equilibrium model with pervasive rational inattention, but they too allocate each variable to a unique agent type as a choice variable. Because this allocation is apparently somewhat arbitrary, they examine variants of their model with different allocations.

Such models are reasonable starts on the project of introducing rational inattention into equilibrium models, but probably we need to go further. In many markets, for example any with continuous trading among many buyers and sellers, the allocation of a price variable to one type of agent as a choice variable does not make sense. We instead see special institutions or types of market participants that allow markets to function without infinite attention from participants. We have retailers, wholesalers, market-makers, and inventories, for example. Recently we have had in asset markets specialist high-frequency traders that process market information at a high rate, using powerful computers. Conventional economic theory, with all agents continuously optimizing using all available information, finds it difficult to explain the role of these specialized economic roles and institutions. At this point, rational inattention has not provided any theory for these institutions and roles either, but it seems to be a promising starting point for such a theory.

Another issue that arises in bringing rational inattention to equilibrium models is that the RI models of individual behavior have nothing to say about properties of information processing error other than its conditional distribution given decision choices. Consider commuters who regularly drive past several gas stations on the way to work. They might not usually pay much attention to gas prices,

stopping at stations randomly, or at some customary station, but if one station cut prices sharply, they might, after a day or two, notice and take advantage of the low price. Which day they noticed might be random and uncorrelated across the commuters. On the other hand, many of them might talk to each other, or the local newspaper might run a story on the unusual behavior of gas prices, in which case the timing of their reaction to the price, while no less "noisy", might be highly correlated across commuters. Information-processing noise that is independent across agents will average out in macroeconomic behavior, whereas highly correlated information processing noise will become an additional source of macroeconomic randomness.

4. Implications for macroeconomic modeling

4.1. **Be more relaxed about micro-foundations for dynamics.** Rational inattention models are difficult to work with and there remain serious substantive issues about how to formulate such models as equilibrium systems. Nonetheless, from the kinds of qualitative results we have described in previous sections, there are some important implications for modeling practice. Rational inattention is a potential explanation for much of the inertia we see in economic behavior, yet its implications are in many respects quite different from those of other hypotheses about the sources of inertia. This suggests that for the time being it may not be a good idea to insist on specific microeconomic stories about the sources of inertia. Invocation of rational expectations micro-theory models to justify constraints on model dynamics may be a mistake. Use of such microeconomic stories to justify welfare evaluations of alternative policies may also be a mistake. On the other hand, resorting to models that

pay no attention to the pervasive inertia and noisiness that we actually observe in dynamic economic behavior would be an even bigger mistake.

We should recognize that many aspects of economic behavior will show slow and erratic adjustment in the direction predicted by optimizing theory, without insisting that agents react as quickly and precisely as rational expectations dynamics would suggest. A promising route forward in this respect is represented by the work of DelNegro, Schorfheide, Smets, and Wouters (2007). They lay out a method for using a rational expectations equilibrium model to generate a prior distribution for the form of a structural vector autoregression (SVAR). The SVAR is left free to match the dynamics in the actual data, to the extent that the data has a strong message about the dynamics, while aspects of the model about which the data do not speak strongly conform to the equilibrium model. Since data generally have much weaker information about long run than about short run dynamics, this has the effect of putting emphasis on the equilibrium model for the long run, and on the data fro the short run. Their method could arguably be improved footnoteSee my comments on the paper in the same issue of the journal., but it seems a step in the right direction and has already been widely applied.

4.2. **Local expansions?** Most of the work in economics that applies rational inattention has focused on the linear-quadratic Gaussian case. This fits well with the fact that most of the use of economic equilibrium models fitted to data has entailed working with their local expansions, often just linear expansions, about deterministic steady states. There is a reason for caution, here, however. Working with low-order local expansions of a nonlinear equilibrium model is justified under reasonable regularity conditions when the initial conditions are close to the steady

state and the scale of disturbance variation is small.[6] But in models with a fixed cost of information, like (4)-(5) above, as we let the scale of random variation in the disturbances shrink, information collection goes to zero before disturbances have gone to zero. That is, there is generally a level of random variation so small it is optimal for no information at all to be collected.

This paradox does not arise if the problem is formulated with fixed Shannon capacity rather than a fixed cost of information processing. As we have already argued, though, it is more appealing to think of people as applying a small part of their full information processing capacity to monitoring economic signals, with a stable shadow price on that processing capacity, than to suppose that they have a fixed capacity constraint.

In order to end up with a model that is well approximated as linear-quadratic and Gaussian we must think of the scale of economic disturbances to the model as "small", and at the same time think of the shadow price of Shannon capacity as small. As documented in every application of rational inattention, to get interesting and realistic effects on dynamics requires that information about individual economic variables be processed at a rate of a few bits per month or quarter. Variations in processing rate in that range probably are realistically modeled as having a stable opportunity cost to individuals.

It might seem that the fact that, as we discussed in section 3.4, optimal behavior of capacity-limited agents often implies discrete behavior would undermine the validity of local LQ Gaussian expansions. This is not necessarily true, however. While it is true that, with initial uncertainty truncated-Gaussian and a quadratic

[6]See Kim, Kim, Schaumburg, and Sims (2008) for one such set of conditions.

loss function, behavior will emerge as discretely distributed, the number of points in the discrete distribution grows larger as the truncation points become larger in absolute value relative to the standard deviation of the initial uncertainty. The discretely distributed behavior becomes distributed over a finely spaced grid of many points, and its distribution becomes close in the metric of convergence in distribution to a Gaussian distribution, despite remaining discrete.

Though we have presented no formal argument proving this, it does seem then that using local linear expansions of models with rational inattention and maintaining Gaussian assumptions on randomness can be justified. But the conditions that justify this should be kept in mind. In periods of economic disruption — hyperinflations or financial crises, for example — stochastic disturbances are large and people may in fact devote a large fraction of their information-processing capacity to tracking economic signals. In some markets, particularly financial markets, there are some people whose full time job consists of tracking price signals and making trades. Those people's behavior, and hence those markets, are therefore probably not well approximated by linear-quadratic Gaussian rational inattention models, though implications of rational inattention may be even more important in studying the short term dynamics of such markets than in most macroeconomic applications. At the other extreme, we should bear in mind that it is possible for optimal behavior to imply ignoring variation in some economic signals because the information costs of attending to it at all do not justify the returns from doing so.

5. IMPLICATIONS FOR MONETARY POLICY

5.1. **A critique of rational expectations policy evaluation.** One of the main insights about policy from rational expectations theory has been the "rational expectations critique of econometric policy evaluation". This is the point that, because the stochastic process followed by the economy changes when macroeconomic policy changes, private sector agents' forecasting rules also change with economic policy. This implies that to project the long run effects of a policy change, one must calculate the change induced in the stochastic process, accounting for the fact that private sector forecasting rules change.

But in a standard rational expectations model agents forecast optimally, no matter how small or smooth are stochastic fluctuations in the economy. Agents respond to optimal forecasts with the same coefficients, regardless of whether the forecasts are oscillating strongly or are hardly changing.[7] Agents with rational inattention, though, will respond with more delay and information-processing error — or may not respond at all — to fluctuations that are small and therefore relatively unimportant to them. This implies that rational expectations models estimated from periods of stability will imply large adjustment costs, and that these models are then likely to be unreliable in tracking behavior when policy or exogenous shocks become much more volatile.

[7] Strictly speaking, this is true only in a linear or linearized rational expectations model, but the point that coefficients do not shrink when shocks become small in a rational expectations model, while they do shrink as shocks become small in a rational inattention model, remains valid.

There is in other words a "rational inattention critique of rational expectations policy evaluation". The rational expectations critique of econometric policy evaluation has sometimes been interpreted to mean that use of econometric model conditional forecasts in policy formation is pointless or misleading, as this sort of exercise seldom accounts explicitly for endogenous shifts in expectation-formation in reaction to changed policy rules. As I have argued elsewhere (1987), this is a mistake. Most real time policy-making is is the non-trivial task of implementing a policy rule that changes little if at all. A correctly identified model can make useful conditional projections of the effects of policy choices without separately identifying the part of its effects that arise from shifts in expectation-formation rules. On the other hand, when we contemplate major changes in policy, we should keep in mind possible rational expectations effects on forecasting rules.

These same points apply to rational inattention. Usually, the effects of rational inattention on the economy's dynamics take a stable form, so that we can project the effects of policy actions without an explicit model of how rational inattention affects those dynamics. But when there are major shifts in policy or in the nature of exogenous disturbances, we should keep in mind that apparent inertia in historical data from less turbulent times could change character as people shift their attention.

5.2. **Monetary policy transparency.** Central bankers sometimes have the impression that financial markets and the press misinterpret or overintepret their policy announcements. The US Federal Reserve makes brief written policy statements after each of the periodic open market committee meetings. The wording of these statements sometimes changes only slightly from one meeting to the next, and the

changes in wording are the subject of close analysis by financial market participants and the press. This is sometimes seen as a reason for being parsimonious about handing out information. If small amounts of information produce overreactions in financial markets, after all, wouldn't large amounts of information produce even worse overreactions? And if sophisticated financial experts misinterpret information, wouldn't increased transparency produce even worse misinterpretation by the general public?

A rational inattention perspective suggests that this reasoning has it backwards. Financial market participants are likely to attend to every nuance of whatever information about its policy that the central bank supplies. If the central bank supplies little information, financial experts will make their own estimates of what lies behind the policy statements and will inevitably make some mistakes. Ordinary people will most likely pay little attention to even simple policy announcements, and they will react sluggishly — in effect simplifying the policy statement through their own information-processing filters — whether the information supplied is dense and complex or simple. This might suggest that there is no harm in simply providing detailed information about policy, and as a first approximation this is indeed what rational inattention theory would suggest. Once we recognize, though, that it is inevitable that complex information will be perceived by the public with delay and error, there is an argument for guiding the simplification of the policy message. In effect, by providing its own simplified summary of a more detailed description of policy, the central bank can do some of the work of "coding" the policy statement into a form that the public can track more directly.

Most inflation-targeting banks provide policy statements called inflation reports at regular intervals, and these often have a two-tiered format. A simple and brief characterization of policy and the state of the economy starts the report, and more detail is provided in later pages. This seems like the right approach: a short, low-information-content summary to guide people who will give the announcement only slight attention, together with detail for those who have reason to read it closely.

Some central banks (e.g. those of New Zealand, Norway and Sweden) have begun providing information about expected future time paths of policy rates. One argument against this practice has been that it could undermine central bank credibility. The public might focus on, say, a projected interest rate one year ahead, and become disillusioned when, inevitably, the forecast turned out to be inaccurate. But central banks that have taken this course have done so in the context of detailed, regularly updated, inflation reports, of which interest rate forecasts are only one element, and often not the most newsworthy element. Interest rate forecasts are usually displayed as "fan charts" that inhibit their interpretation as simple numerical targets. Since people are unlikely to have loss functions that make minor deviations of forecast from actual interest rates important to them, they are unlikely to focus narrow attention on interest rate point forecasts when these are just one part of a richer presentation of information.

6. Directions for progress

We have by now examples of research applying Shannon information-theoretic ideas in a number of directions in economics and finance. One of the earliest was

Woodford (2002), which cited rational inattention theory as motivation for considering a model in which agents perceive the state of the economy imprecisely. In later work 2009 Woodford uses information theory more formally, while combining it with other sources of inertia. In finance, Mondria (2005), Van Nieuwerburgh and Veldkamp (2004), and Peng and Xiong (2005), for example, have applied information-theoretic ideas. We have already noted the work of Luo (2008) and and Matějka (2009, 2008); Matějka and Sims (2009). Luo and Eric Young have a series of papers that apply a rational inattention permanent income framework to, among other things, asset pricing and the current account, a recent example being Luo, Nie, and Young (2010). Maćkowiak and Wiederholt (2009b,a) have worked out a partial equilibrium model of producers pricing in response to multiple sources of cost variation and, later, a complete dynamic stochastic general equilibrium model in which interacting agents of different types face information processing constraints.

All of these papers are worthwhile efforts, but all make compromises to keep the modeling problem tractable. Only the Mondria paper and my early paper (2003) consider models with a multivariate state variable and recognize the point made in section 3.2.3 that rational inattention induces ex post correlation of uncertainty across initially independent state variables. Some deal with problems in which the state is one-dimensional, while others, like those of Peng and Xiong, van Niewerburgh and Veldkamp, and Maćkowiak and Wiederholt, impose ex post independence on initially independent states as a matter of convenience. In their 2009a paper, Maćkowiak and Wiederholt recognize this limitation on their approach, and try to allow for it by experimenting with what amounts to rotations of the state space. In a multivariate problem, ex post correlation is induced by the fact that

agents will want to collect information only about certain dimensions of variation in the state. By reducing uncertainty in those dimensions, they induce correlation of remaining uncertainty in other dimensions. But if the state vector can be redefined via a linear transformation so that the components about which agents do not collect information are distinct "state variables", there will be no induced ex post correlation. Maćkowiak and Wiederholt's approach is therefore a step in the right direction, though there is no way within their framework to verify that they have checked all relevant rotations of the state vector.

As we have already noted, competitive markets, in which prices are equilibrium phenomena not controlled by any one optimizing agent, raise difficult issues for rational inattention modeling. In macro models, in which it has become conventional to postulate prices set by monopolistically competitive firms, this is not directly an issue. But in finance models, where asset prices are not realistically treated as set by monopolists, it is a serious difficulty. The most interesting models would involve market participants who see the market price only via a capacity-limited channel, but if all agents are so limited, the usual competitive market-clearing mechanisms are not available. Finance models that have attempted to model market equilibrium, like Mondria's, have therefore tended to make schizophrenic compromises, assuming that some external signals (e.g. market prices) are perceived without error, while others are subject to a capacity constraint.

Recent instabilities in asset markets and their macroeconomic consequences have generated renewed interest by economists in trying to understand liquidity. Gorton and Metrick (2009) provide suggestive evidence that economizing on information-processing requirements created demand for some types of securities before the

crash, and the loss in liquidity of these securities as their information-processing requirements increased was a major source of disruption during the crash. It seems likely that insights from information theory can help us understand these phenomena, and there are economists working in this direction, though not with any citable research output to this point.

In modeling asset markets particularly, moving beyond the linear-quadratic Gaussian framework seems important. Even if risky assets have yields with Gaussian distributions, the optimal portfolio problem in the presence of risk aversion is not linear-quadratic, and apparently has not yet been solved, even numerically, under a rational inattention assumption. The result will not be ex post Gaussian uncertainty about yields, and the nature of the induced non-Gaussianity would be interesting to explore. My own work on the two-period savings problem Sims (2005, 2006) and Matějka's previously cited work focus primarily on two-period problems. Matejka considers a very simple dynamic problem. Tutino (2009) takes up a fully dynamic savings problem without assuming normality, but is constrained by computational considerations to work within a fairly small, discrete probability space. Much, therefore, remains to be done in this area.

7. Conclusion

Rational inattention has cast a critical light on much existing financial and macroeconomic modeling, suggesting that the now-standard technical apparatus of rational expectations could easily give misleading conclusions. At the same time, formally incorporating rational inattention into macroeconomic and financial models is an immense technical challenge. While the modest progress to date on these technical challenges may be discouraging, we might take comfort in the fact that

rational expectations itself was seen as imposing immense technical challenges at the outset, so that it took decades for it to become a regular part of policy modeling.

APPENDIX A. GENERAL LINEAR-QUADRATIC CONTROL WITH AN INFORMATION COST

Consider the problem

$$\max_{X_t, \hat{Y}_t, \Sigma_t} E\left[\sum_{t=0}^{\infty} \beta^t (Y_t' A Y_t + Y_t' B X_t + X_t' C X_t - \lambda H_t)\right] \tag{16}$$

subject to

$$Y_{t+1} = G_1 Y_t + G_2 X_t + \varepsilon_{t+1} \tag{17}$$

$$H_t = \tfrac{1}{2}(\log|M_t| - \log|\Sigma_t|) \tag{18}$$

$$M_{t+1} = \Omega + G_1 \Sigma_t G_1' \tag{19}$$

$$\varepsilon_t \mid \{Y_s, X_s, s < t\} \sim N(0, \Omega) \tag{20}$$

$$M_t - \Sigma_t \text{ positive semi-definite} \tag{21}$$

$$Y_t \mid \mathcal{I}_t \sim N(\hat{Y}_t, \Sigma_t) \tag{22}$$

$$\{X_t, X_{t-1}, \dots\} \subset \mathcal{I}_t\,. \tag{23}$$

Then by the law of iterated expectations we can rewrite the objective function as

$$E\left[\sum_{t=0}^{\infty} \beta^t (\text{trace}(\Sigma_t A) + \hat{Y}_t' A \hat{Y}_t + \hat{Y}_t' B X_t + X_t' C X_t - \lambda H_t)\right], \tag{24}$$

where $\hat{Y}_t$ is $E[Y_t \mid \{X_t, X_{t-1}, \dots\}]$. Since H_t depends on Σ_t and Σ_{t-1}, but not on any values of X or $\hat{Y}$, the objective function is the sum of two pieces, one a function of only the X and $\hat{Y}$ values, the other depending only on Σ_t and M_0.

We can also rewrite the dynamic constraint (17) as a constraint in terms of $\hat{Y}$:

$$\hat{Y}_{t+1} = G_1 \hat{Y}_t + G_2 X_t + \xi_{t+1} \tag{25}$$

$$\text{with } \xi_t = \hat{Y}_t - Y_t + G_1(Y_{t-1} - \hat{Y}_{t-1}) + \varepsilon_t \,. \tag{26}$$

The error term ξ_t in this equation has two components in addition to the original disturbance ε_t, both of which are uncorrelated with any element of $\mathcal{I}_{t-1}$. The first, $\hat{Y}_t - Y_t$ is minus the error of prediction of Y_t based on the larger information set $\mathcal{I}_t$, and is therefore uncorrelated with anything in $\mathcal{I}_{t-1}$. The second is a linear function of the error in the best predictor of Y_{t-1} based on $\mathcal{I}_{t-1}$, and is therefore also uncorrelated with anything in $\mathcal{I}_{t-1}$. Thus the problem has as one component a conventional linear-quadratic stochastic control problem:

$$\max_{X_t, \hat{Y}_t} E\left[\sum_{t=0}^{\infty} \beta^t (\hat{Y}_t' A \hat{Y}_t + \hat{Y}_t' B X_t + X_t' C X_t\right]$$

subject to (25). This can be solved for the optimal linear relation between X_t and Y_t using certainty equivalence, since the variances of disturbances do not affect the solution.

While the solution of the embedded linear quadratic control problem does not depend on the disturbance variances, the value function for the problem does, in general. We will not try to present a general solution method here. However in the examples considered in this paper, because they are "tracking problems", the value function for the linear quadratic problem is trivial. The optimal certainty-equivalent solution makes X and Y match perfectly and delivers zero losses. Thus the terms in the objective function involving $\hat{Y}$ and X drop out, leaving the the

deterministic problem

$$\max_{\Sigma_t} \sum_{t=0}^{\infty} \beta^t (\text{trace}(\Sigma_t A) - \lambda H_t) \tag{27}$$

subject to

$$H_t = \tfrac{1}{2}(\log |M_t| - \log |\Sigma_t|) \tag{28}$$

$$M_t = \Omega + G_1 \Sigma_{t-1} G_1' \tag{29}$$

$$M_t - \Sigma_t \text{ positive semi-definite.} \tag{30}$$

For this Σ_t part of the problem, the first order condition, if we ignore the positive-definiteness constraint (30), is

$$A = \beta\lambda G_1 M_{t+1}^{-1} G_1' - \lambda \Sigma_t^{-1} \,. \tag{31}$$

If the positive-definiteness constraint does not bind, this is (after using (29) to eliminate M_{t+1}) a nonlinear equation in Σ_t that can be solved by standard methods. A starting point for a solution, therefore, will generally be to solve this equation and check whether in fact (30) is satisfied by the solution value of Σ_t and the initial M_0. If so, the problem is solved. If not, in the univariate case, the solution is still straightforward, because the model is implying that even when no information is collected, so X_t is just a constant, the contribution of additional information is less than its cost. It is possible that with no information collected M_t will grow to the point where it exceeds the optimal Σ_t, after which Σ_t remains constant at its optimal value. In the general case, though, we have to treat the solution for Σ_t as a constrained nonlinear deterministic dynamic programming problem.

Even in the simple two-dimensional tracking problem of section 3.2.4, the positive-definiteness constraint binds. The problem can be solved by making the Cholesky decomposition of $M_t - \Sigma_t$ the solution parameter, using a Cholesky decomposition constrained to be of a fixed, less than full, rank, and applying the chain rule to convert the first order conditions with respect to Σ in (31) to FOC's with respect to the new parameters.

Note some implications of this general treatment. In tracking problems in which information enters the objective function with a fixed cost per bit, the optimal solution will eventually imply a constant Σ_t. That is, the uncertainty about the state will not vary with the level of the state variable. Also, when information costs are low enough and initial uncertainty large enough, the solution will move immediately to its steady state value. And finally, in a multivariate problem it can happen that $M_t - \Sigma_t$ is only positive semi-definite, not positive definite, implying that information is optimally collected only about certain dimensions of uncertainty about the state vector.

REFERENCES

AKERLOF, G. A., AND J. L. YELLEN (1985): "Can Small Deviations from Rationality Make Significant Differences to Economic Equilibria?," *The American Economic Review*, 75(4), 708–720.

BIERBRAUER, J. (2005): *Introduction to Coding Theory*, Discrete Mathematics and Its Applications. Chapman and Hall/CRC.

COVER, T. M., AND J. A. THOMAS (1991): *Elements of Information Theory*. Wiley-Interscience.

DELNEGRO, M., F. SCHORFHEIDE, F. SMETS, AND R. WOUTERS (2007): "On the Fit and Forecasting Performance of New Keynesian Models," *Journal of Business and Economic Statistics*, 25(2), 123–162.

EICHENBAUM, M., N. JAIMOVICH, AND S. REBELO (2008): "Reference Prices and Nominal Rigidities," Discussion paper, Northwestern University and Stanford University, NBER Working paper 13829.

GORTON, G. B., AND A. METRICK (2009): "Securitized Banking and the Run on Repo," Working Paper 15223, National Bureau of Economic Research, http://www.nber.org/papers/w15223.

KIM, J., S. KIM, E. SCHAUMBURG, AND C. SIMS (2008): "Calculating and Using Second Order Accurate Solutions of Discrete Time Dynamic Equilibrium Models," *Journal of Economic Dynamics and Control*, 32(11), 3397–3414.

LUO, Y. (2008): "Consumption dynamics under information processing constraints," *Review of Economic Dynamics*, 11(2), 366 – 385.

LUO, Y., J. NIE, AND E. R. YOUNG (2010): "Robustness, Information-Processing Constraints, and the Current Account in Small Open Economies," Discussion paper, University of Hong Kong, http://yluo.weebly.com/uploads/3/2/1/4/3214259/carbri2010h.pdf.

MACKAY, D. J. C. (2003): *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

MAĆKOWIAK, B., AND M. WIEDERHOLT (2009a): "Business Cycle Dynamics under Rational Inattention," Discussion paper, European Central Bank and Northwestern University, http://faculty.wcas.northwestern.edu/ mwi774/RationalInattentionDSGE.pdf.

——— (2009b): "Optimal Sticky Prices under Rational Inattention," *American Economic Review*, 99(3), 769–803.

MANKIW, N. G., AND R. REIS (2002): "Sticky Information Versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve*," *Quarterly Journal of Economics*, 117(4), 1295–1328.

——— (2010): "Imperfect Information and Aggregate Supply," in *Handbook of Monetary Economics*. Elsevier.

MATĚJKA, F. (2008): "Rationally Inattentive Seller: Sales and Discrete Pricing," Discussion paper, PACM, Princeton University, http://www.pacm.princeton.edu/publications/Matejka_F_2008-wp.pdf.

——— (2009): "Rigid Pricing and Rationally Inattentive Consumer," Discussion paper, Princeton University.

MATĚJKA, F., AND C. SIMS (2009): "Discrete actions in information-constrained tracking problems," Discussion paper, Princeton University.

MONDRIA, J. (2005): "Financial Contagion and Attention Allocation," Discussion paper, Princeton University.

PENG, L., AND W. XIONG (2005): "Investor Attention, Overconfidence and Category Learning," Discussion paper, Princeton University.

SIMS, C. A. (1987): "A rational expectations framework for short-run policy analysis," in *New approaches to monetary economics*, ed. by W. A. Barnett, and K. J. Singleton, pp. 293–308. Cambridge University Press, Cambridge, England.

——— (1998): "Stickiness," *Carnegie-rochester Conference Series On Public Policy*, 49(1), 317–356.

——— (2003): "Implications of Rational Inattention," *Journal of Monetary Economics*, 50(3), 665–690.

——— (2005): "Rational Inattention: A Research Agenda," Discussion paper, Princeton University, http://sims.princeton.edu/yftp/RIplus, http://sims.princeton.edu/yftp/RIplus/.

——— (2006): "Rational Inattention: Beyond the Linear-Quadratic Case," *American Economic Review*, 96(2), 158–163.

TUTINO, A. (2009): "The Rigidity of Choice: Lifetime Savings under Information-Processing Constraints," Ph.D. thesis, Princeton University, http://docs.google.com/fileview?id=0B7CdO9AORsjcNWYwZmM1MWEtNDZiNi00NzQzLTgzOTItZmNiM2IzOWQ3MDhh&hl=en.

VAN NIEUWERBURGH, S., AND L. VELDKAMP (2004): "Information Acquisition and Portfolio Under-Diversication," Discussion paper, Stern School of Business, NYU.

WOODFORD, M. (2002): "Imperfect Common Knowledge and the Effects of Monetary Policy," in *Published in, eds., Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, ed. by P. Aghion, R. Frydman, J. Stiglitz, and M. Woodford. Princeton University Press, http://www.columbia.edu/ mw2230/phelps-web.pdf.

——— (2009): "Information-Constrained State-Dependent Pricing," *Journal of Monetary Economics 56(S): 100-124 (2009)*, 56(S), 100–124.