Stickiness

October 20, 2020

Why does it matter?

- Keynes's idea: If wages don't adjust fast enough to clear markets, while firms still act like "price-takers" in the labor market, then nominal wealth changes (aggregate eemand) and monetary policy changes that otherwise would only change the price level, have real effects.

- New Keynesian models: firms set prices in monopolistic competition, but face some sort of cost to change, or to rapidly change, prices.

Costs of adjustment

- Calvo
- Rotemberg
- Mankiw-Reis

Continuous time math

$$\min \int_0^\infty e^{-\beta t}\big((p_t - p_t^*)^2 + \gamma \dot{p}_t^2\big)$$

FOC:

$$2(p_t - p_t^*) + 2\beta\gamma\dot{p}_t - 2\gamma\ddot{p} = 0$$
$$-(\gamma D^2 + \beta\gamma D + 1)p_t = p_t^*$$

One root positive, one negative, both real. The positive root solves forward.

Solving forward

$$
\begin{aligned}
-\gamma(D-r_1)(D+r_2)p_t &= p_t^* \\
(D+r_2)p_t &= -\gamma^{-1}(D-r_1)^{-1}p_t^* \\
\dot{p}_t + r_2 p_t &= \frac{1}{\gamma}\int_0^\infty p_{t+s}^* e^{-r_1 s}\,ds\;.
\end{aligned}
$$

The right-hand side is a discounted sum of expected future values. It can jump around. If p_t^* is stochastic and continuously varying, this kind of expression has non-differentiable time paths. But it's only $\dot{p}$, not p itself, that jumps around. This is intuitive: If there's a penalty on large $\dot{p}$, jumps or infinite variance derivatives of p are infinitely costly. Adjustment costs imply smooth time paths.

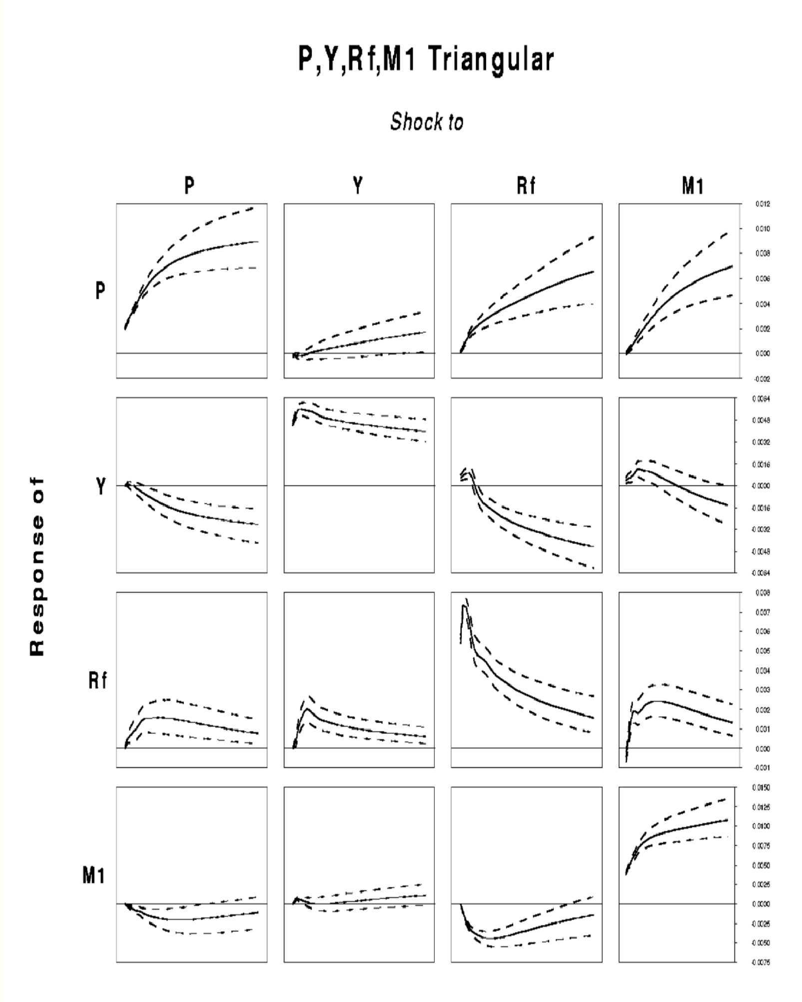
Behavior of p_t paths with menu costs or Calvo

- For individual firms, prices stay constant for a while, then jump.
- With Calvo, aggregate price paths are differentiable, because infinitesimally small fraction of price setters are changing prices at any given time.
- With menu costs, behavior depends on the distribution of prices across firms, relative to p^*.
- Also on relative variances of micro-disturbances to p^* vs. macro-disturbances. (Nakamura-Steinsson, Bils-Klenow).

Summary

- Calvo, Rotemberg, menu costs in quiet times, imply smooth aggregate time paths.
- Menu costs imply that jumpy micro-behavior due to high local variability in p^* should make responses to macro shocks fast also.
- All of this theory, and the recent literature looking at micro price data, focus on *price* stickiness.
- They, and also Mankiw-Reis, imply that when prices do change at the micro level, they respond to all information available to anyone in the economy.

Aggregate facts



Facts

- For prices *and quantities* time paths are not smooth — see the diagonals of the irf plots.

- For prices *and quantities* cross-variable responses are smooth and delayed.

- These patterns don't fit any of the "cost of adjustment" stories.

Weak and strong rational expectations

- Weak: Individual agents optimize continually based on their probabilistic beliefs and information.
- Strong: Everyone has the same probability distribution and information set at every moment — except possibly the government.

Weak and strong rational expectations

- Weak: Individual agents optimize continually based on their probabilistic beliefs and information.
- Strong: Everyone has the same probability distribution and information set at every moment — except possibly the government.
- Lucas "International evidence" paper: People can't fully separate micro and macro sources of disturbance to p^*.

Why did applied macro almost entirely abandon the Lucas interpretation of RE?

- Theory: In that framework, if people can see an interest rate or the money supply with little delay, real effects of nominal shocks entirely, or almost entirely, disappear.

- Practice: RE models with agents with different information sets are difficult to handle, and in a sense fragile (Grossman-Stiglitz).

Stickiness from information processing

Introspection:

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

Stickiness from information processing

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.

Stickiness from information processing

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.

Stickiness from information processing

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.

Stickiness from information processing

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.

Stickiness from information processing

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?

Stickiness from information processing

Macroeconomics:

- This way of thinking about stickiness captures the intuition of why "everyone can see M_t" seems artificial as a way to dismiss the Lucas micro/macro confusion idea.

Stickiness from information processing

Macroeconomics:

- This way of thinking about stickiness captures the intuition of why "everyone can see M_t" seems artificial as a way to dismiss the Lucas micro/macro confusion idea.

- Just because eeryone can *see* $M1$ doesn't mean everyone *pays attention* to $M1$.

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.

Preview of qualitative results from rational inattention

- Decision variables are likely both to have a non-persistent "noise" component arising from imprecise information, and to respond smoothly and with a delay to arriving information.
- This fits the pattern we see in the aggregate time series data.
- Decision variables can easily end up with distributions with finite support, meaning that their values can repeat and that they jump among a finite set of values.
- When this pattern of discrete support emerges, the frequency of the jumps has no necessary connection to the amount of delay in reaction to new information, and agents generally have no better information about the state of the economy at the time of a jump than at other times.

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".

Or, more accurate information observable with a cost

$$\max_{\delta} E[U(Y,X,Z,\delta)] - \theta/\sigma^2 \quad \text{subject to}$$

$$\delta \text{ a function of } Z$$

Ideally, we would like to make δ a function of X

$$\text{e.g. } U(Y,X,Z,\delta) = -(X - \delta(Z))^2$$

$$\mathrm{Var}(X \mid Z) = \sigma^2 \,.$$

Here θ is one form of a "cost of information".

Applicability of the traditional approach

- Where there is a physical cost of obtaining information — running a survey, drilling a test well, hiring an RA — the traditional approach applies well.

- It leaves the form of the costs unstructured, except by a concrete application.

- Should costs be linear in variance, one over variance, standard deviation rather than variance, interquartile range?

- The approach is not helpful in thinking about human information *processing* costs, where no resource use outside the decision maker's brain is involved.

Example of a physical cost: repeated sampling

- We can run a survey of n individuals. Cost might be close to linear in n.
- Variance σ^2 is likely to be proportional to $1/n$.
- Then information cost is proportional to $1/\sigma^2$.
- The Shannon measure we are about to discuss does not behave this way.

The channel

- This is Shannon's seminal abstraction.

The channel

- This is Shannon's seminal abstraction.
- It has an input **alphabet**, which is a list of possible inputs to the channel.
- It describes the conditional distribution of an output alphabet given any element of the input alphabet.

The channel

- This is Shannon's seminal abstraction.
- It has an input **alphabet**, which is a list of possible inputs to the channel.
- It describes the conditional distribution of an output alphabet given any element of the input alphabet.
- That's it.

Examples of channels

- Simplest: Telegraph key. Input alphabet: "dit", "dah". Output alphabet: same. Could be error free, or with some probability of error.

- Input alphabet: musical tones. Output alphabet: musical tones, but with some noise.

- Input alphabet: real numbers, but with a bound on their variance. Output alphabet, real numbers, with known conditional distribution given input.

Shannon's question

- You have a message to send and a channel. Alphabet of the message does not match the input alphabet of the channel. E.g.: message is Shakespeare's *The Tempest*; channel is a telegraph key with some error.

- This means you have to *encode* your message. For example, Morse code.

- What is the most efficient way to encode it, so as to send it as fast as possible with minimal error?

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.

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 $H(p_1, \ldots, p_n)$

Additivity

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:

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

Uniqueness

The additivity property is enough to tell us that $H(Y) = H(p_1, \ldots, p_n) = -E[\log(p_i)]$. We can add 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.

Example

Four points in the probability space, indexed by $i \in \{1, 2, 3, 4\}$, all with equal probability. $Y(i) = i$. $X_1(i) = 10 \cdot \mathbb{1}[i \leq 2]$, $X_2 = 3 \cdot \mathbb{1}[i = 1]$, $X_3 = 0.8 \cdot \mathbb{1}[i = 4]$.

$$H(X_1) = 1 \tag{1}$$

$$H(X_2 \mid X_1) = .5 \times 1 + .5 * 0 = .5 \tag{2}$$

$$H(X_3 \mid X_2, X_1) = .5 \times 0 + .5 \times 1 = .5 \tag{3}$$

$$H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_2, X_1) = H(Y) = 2 \tag{4}$$

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 $-\int_{\mathbb{R}} p(x) \log(p(x))\, dx$.

Learning the exact value of a random variable

- In the discrete case, entropy is reduced to zero.
- In the continuous case, entropy becomes $-\infty$.

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 $H(X \mid Y) = 0$. For the discrete case, this occurs if and only if there is no uncertainty about X after we observe Y.

Invariance under monotone transformations

- $I(X,Y) = I(f(X), g(Y))$ for monotone functions f and g.
- Entropy does not share this property for continuous variables.

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

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

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

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

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

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

FOC

$$
\begin{aligned}
&U(X, Y) \\
&= \theta \left(1 + \log f(x,y) - 1 - \log\left(\int f(x,y')\, d\mu(y')\right)\right) - \psi(y) && (5) \\
&\text{or} && (6) \\
&q(y \mid x) = e^{\left(\psi(y) + U(x,y)\right)/\theta}\,. && (7)
\end{aligned}
$$

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

When the $f(x,y) \geq 0$ constraint does not bind anywhere, the solution is often fairly easy to find, even analytically. But often it does bind, and the solution becomes much harder.

Sort of generic discreteness of solutions

- When the support of g is bounded, x is one-dimensional, and U is $-\left\|(Y-X)^2\right\|$, the optimal dsitribution of X has finitely many points of support. This generalizes to U's that are analytic functions of $Y-X$ under some mild additional regularity conditions.

- See figure for example of sharp difference in optimal X distributions, with little difference in objective function.

CAPM with RI

- See table and figures in JKMS.

LQ Gaussian case

- The FOC's require that

$$\int e^{(\psi(y)+U(x,y))/\theta}\,dy = 1 \tag{8}$$

(since this is the density $q(y \mid x)$, which of course integrates to one).

- If $U(x,y) = (x-y)^2$, we can choose $\psi(y)$ to be constant. That makes the form of $q(y \mid x)$ in the FOC Gaussian, with a variance $\theta/2$.

- This implies that Y has the marginal distribution of X plus a Gaussian random variable ε independent of X. Many, but not all $g(y)$ densities satisfy this condition.

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

- The question is, with $q(y \mid x)$ of the form in (7) with $\psi(\ )$ constant, does the following equation hold?

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

- Some g's satisfy this condition (e.g. a $N(0,\theta)$) or a mixture of two normals, both with variances greater than or equal to $\theta/2$, but some can't (e.g. $U(0,1)$, Gamma, $N(0,\theta/4)$)

- When g is of a form that does not allow satisfying this condition, it must be that $f(x,y) = 0$ in some parts of the x, y space.

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 $\Sigma_Y - \Sigma_{Y|X}$. We are also using the fact that the entropy of a multivariate normal variable with covariance matrix Σ is $\frac{1}{2}\log(|\Sigma|)$ plus a constant.

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.s.d. 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 normalized as square and triangular, 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 ($\Sigma_{Y|X} = \theta I$).