MSc Research Methods: Theory, Text and Matlab

Michael McMahon

University of Warwick

2014-2015

To Cover

- Theoretical research overview
- A description of some theoretical research
  1. micro
  2. econometrics
  3. macro
- Textual analysis
- Matlab introduction

Theoretical Research

Why are you writing a model?

- To prove the existence of an effect:
  - Start simple and then extend/generalise later
- Questions to ask yourself:
  1. Would it be easier to derive the results in continuous or discrete time?
  2. Can I demonstrate the effect in two or three time periods, or do I need to do an infinite horizon model?
  3. Are the functional forms for the utility and production functions the ones that make the derivations as simple as possible?
  4. If the model involves agents of different types, how many different types are actually needed to generate the result?

Theoretical Research

Why are you writing a model?

- Building a theoretical model for empirical implementation:
  - more important to have realistic assumptions
- Questions to ask yourself:
  1. Can you map the model to the data that is available?
  2. Do any simplifications you have to make matter for identification?
  3. Are your instruments good instruments in your model and in the data?

Theoretical Research

Hal Varian's Advice

- Look for ideas in the world, not in the journals.
  - Is the idea interesting?
  - Can you explain it to your mother in plain English?
  - Is the idea important?
- First make your model as simple as possible, then generalize it.
- Look at the literature later, not sooner.
- Model your paper after your seminar.
- Stop when you've made your point.

Tools and Tricks: See the examples

Some related tricks or tools:

- Numerical simulation
- Monte Carlo studies
- Solution techniques to solve forward looking models
  1. Pertubation techniques
  2. Projection methods

Some issues in central bank design

Selection from Reis (2013):

- The strictness of the central bank's mandate
- The choice of long-run goals
- The potential role of additional short-term goals
- The choice of central banker(s)
- The set of assets held by the central bank
- The importance of announcements and commitments
- Choosing the extent of transparency

Increasing use committees of experts

- Central banks don't set interest rates - policymakers set interest rates
- And policymakers are people!

- Pollard (2004) reports that ninety percent of eighty-eight surveyed central banks use committees to decide interest rates
- Committees are groups of policymakers
- Does the appointment of different people matter, and how many people should we have? [Who?]
- Do policymakers behave as theory predicts? [How?]
- What happens behind closed doors in the committee meetings? [How?]

My research papers I will touch on

1. **“How Experts Decide: Preferences or Private Assessments on a Monetary Policy Committee?”** with Stephen Hansen (UPF) and Carlos Velasco Rivera (Princeton)
2. **“Estimating Bayesian Decision Problems with Heterogeneous Priors”** with Stephen Hansen (UPF) and Sorawoot Srisuma (Surrey)
3. **“Transparency and Deliberation within the FOMC: A computational linguistics approach”** with Stephen Hansen (UPF) and Andrea Prat (Columbia)

Model of Individual Decision Making

- C chooses one of two interest rates $r_t \in \{0, 1\}$
- Unknown state of the world: $\omega_t \in \{0, 1\}$
  - Shock drawn from a Bernoulli distribution with $\Pr[\omega_t = 1] = q_t \in (0, 1)$.
- $\pi_t(r_t, \omega_t)$

Member Preferences: θ

- θ is a members preference type
- higher (lower) θ means “hawkish” (“dovish”)
- θ reflects a members “burden of proof” (Feddersen & Pesendorfer (1998)):

		vote	
		0	1
state	0	0	$-(1-\theta)$
	1	$-\theta$	0

Examples of more “macro” preferences

- Could reflect:
  - different inflation targets
  - a different weight on output in a standard loss function
- C’s period t utility is a weighted sum of inflation and output given by

$$\begin{aligned} u_t\left(\omega_t, r_t, \pi_t^P \mid \theta\right) &= W\left[\pi_t(r_t, \omega_t) \mid \theta\right] + \phi(y_t - y^*) \\ &= W\left[\pi_t(r_t, \omega_t) \mid \theta\right] + \chi(\theta^{-1})\left(\pi_t - \pi_t^P\right) \end{aligned}$$

- Higher θ means less evidence needed of high shock to vote high

Uncertainty and Information

- C knows common information q_t
- C observes a private signal s_t
- We assume signals are normally distributed:
  - Private signal $s_{it} \sim N(\omega_t, \sigma_i^2)$
- Best guess of the state of the economy:

$$\ln\left[\frac{\widehat{\omega}_{it}}{1-\widehat{\omega}_{it}}\right] = \ln\left[\frac{q_t}{1-q_t}\right] + \frac{2s_{it}-1}{2\sigma_i^2}. \qquad (1)$$

Decision rule is simple

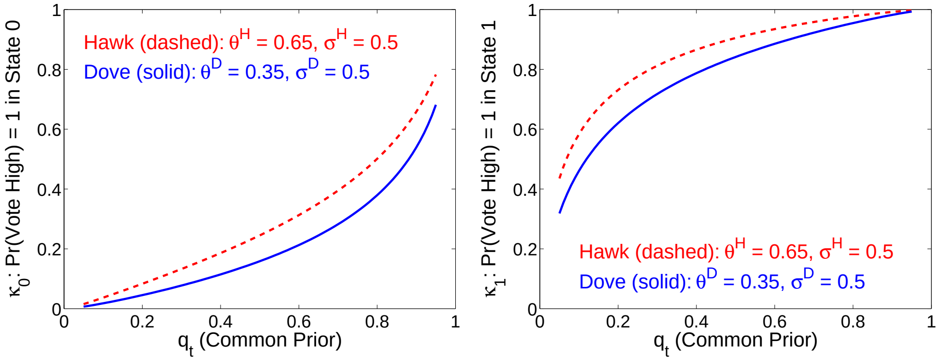
C chooses $v_{it} = 1$ whenever

$$\ln\left[\frac{\widehat{\omega}_{it}}{1-\widehat{\omega}_{it}}\right] \geq \frac{1-\theta_i}{\theta_i} \qquad (2)$$

which implies

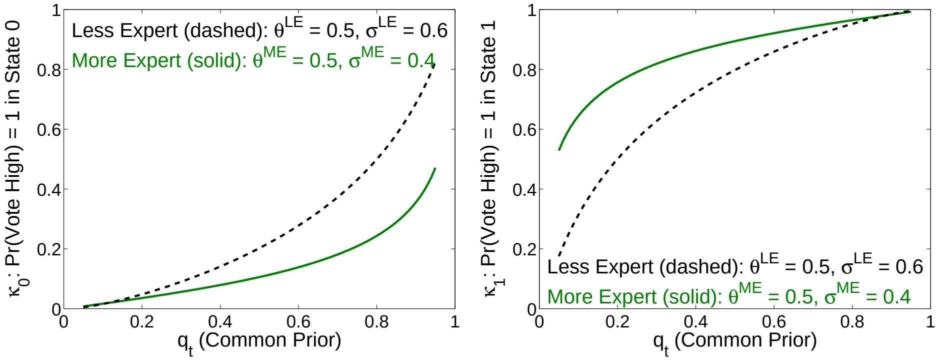
$$s_{it} \geq \frac{1}{2} - \sigma_i^2\left[\ln\left(\frac{\theta_i}{1-\theta_i}\right) + \ln\left(\frac{q_t}{1-q_t}\right)\right] \equiv s_{it}^*(\text{SIN}). \qquad (3)$$

Preference Differences



(a) State=0 (b) State=1

Expertise Differences



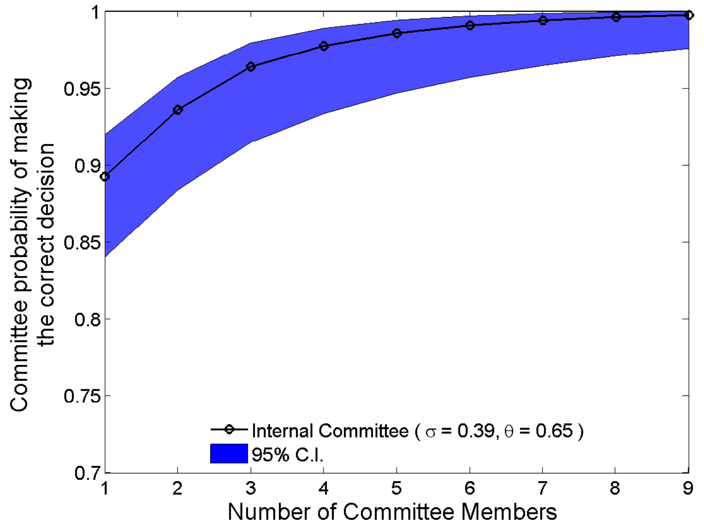
(c) State=0 (d) State=1

Expertise differences

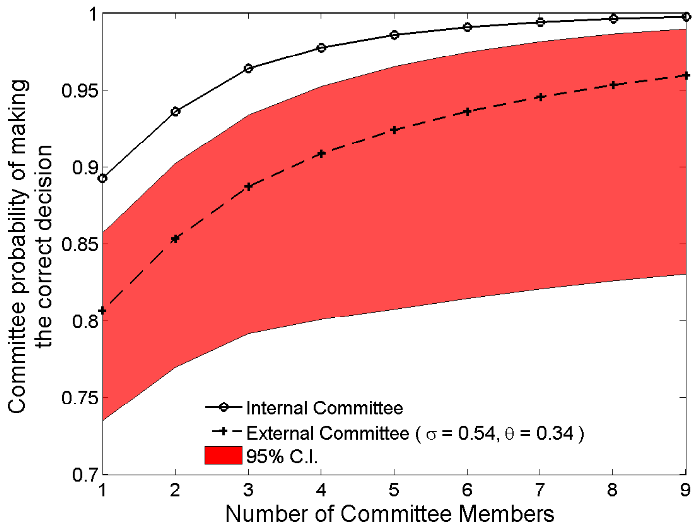
Table: Baseline Estimates of Structural Parameters

	Internal		External		Difference	
σ	0.39		0.54		-0.15	
95% Range	*0.35*	*0.48*	*0.45*	*0.7*	*-0.29*	*-0.04*
θ(SIN)	0.65		0.34		0.30	
95% Range	*0.52*	*0.76*	*0.25*	*0.44*	*0.18*	*0.4*

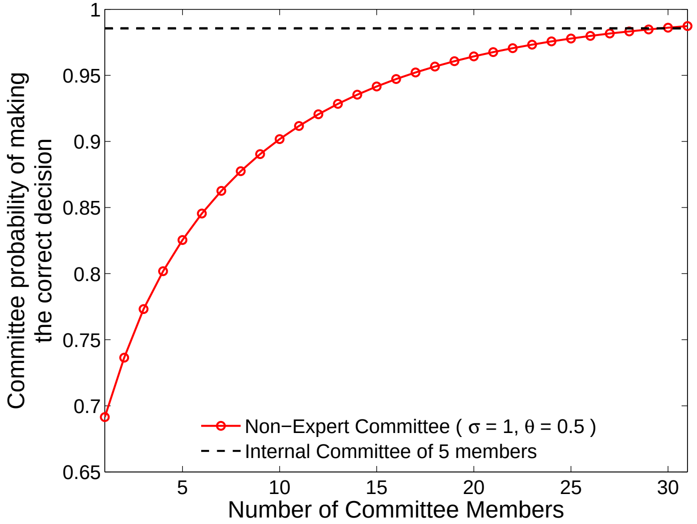
A committee of internal members



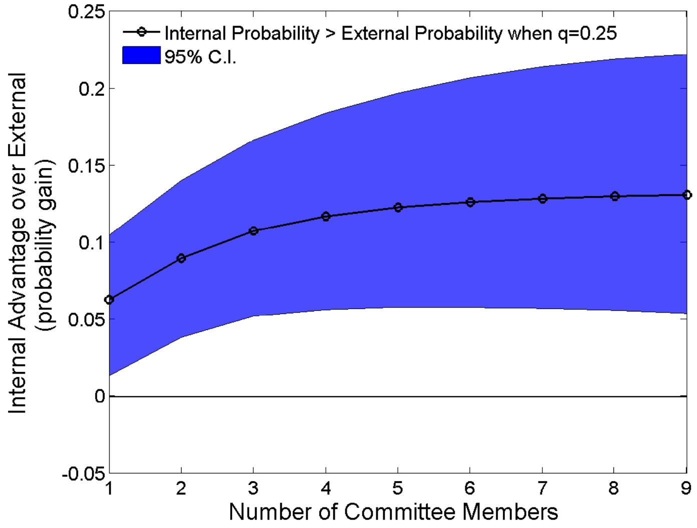
A committee of external members



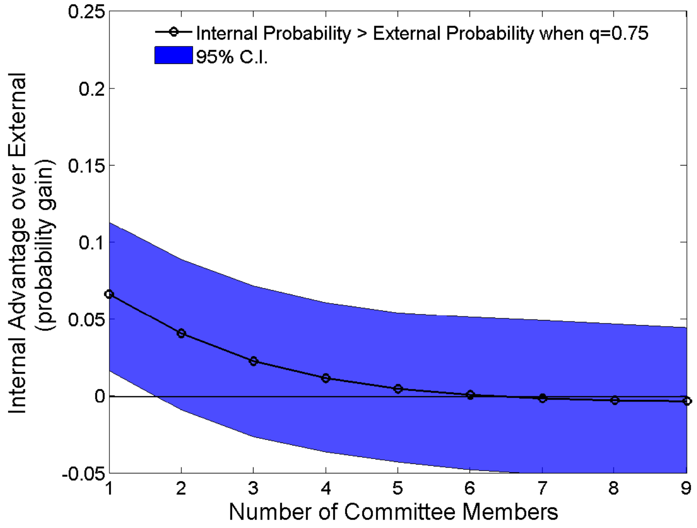
A Non-Expert Committee



Internals advantage over externals: $q_t = 0.25$



Internals advantage over externals: $q_t = 0.75$



Testing our estimator

Two econometric questions of interest

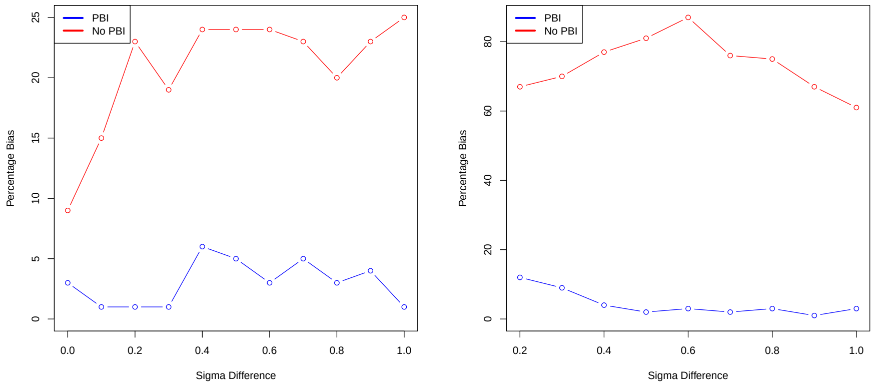
1. Is our estimator accurate?
2. Is our estimator better than existing ones?

Use a Monte Carlo approach to test

Example Monte Carlo

1. Generate a group of 9 decision makers each making 150 decisions in consecutive time periods.
   - 1.1 5 members are type A with preferences θ_A and expertise σ_A; 4 members are type B with preferences θ_B and expertise σ_B
   - 1.2 Use reasonable values ($\theta_A = \frac{2}{3}$, $\theta_B = \frac{1}{3}$, $\sigma_A = 1 - x$ and $\sigma_B = 1 + x$)
2. For each unique set, run 1,000 simulations. For each simulation, we generate theoretical voting data according to the following procedure:
   - 2.1 In each period t, q_t is drawn from $U[0.2, 0.8]$ (independent across periods)
   - 2.2 ω_t is drawn from a Bernoulli distribution with $\Pr[\omega_t = 1] = q_t$
   - 2.3 d_{it} is drawn from a Bernoulli distribution with $\Pr[d_{it} = 1] = 1 - \Phi\left(\frac{s^*_{it} - \omega_t}{\sigma_i}\right)$ where Φ is the normal cdf and s^*_{it} is member i's "critical" threshold.
3. Given these data, we estimate two separate specifications for $\widehat{\theta}_j$ and $\widehat{\sigma}_j$ for $j \in \{A, B\}$

Results of experiment



(e) Bias in $\frac{\theta}{1-\theta}$ Difference Estimates (f) Bias in σ Difference Estimates

Computer RBC

- Production Function

$$Y_t = K_t^{1-\alpha}(A_t N_t)^{\alpha} \qquad 0 < \alpha < 1 \tag{4}$$

- Capital Accumulation

$$K_{t+1} = (1-\delta)K_t + Y_t - C_t \tag{5}$$

- Technological change $A_t = A_t^* \tilde{A}_t$
    - Deterministic Component: $GA_t^* = A_{t+1}^*$
    - Shock process: $a_t = \phi a_{t-1} + \varepsilon_t$ where $a_t = \log \tilde{A}_t$, and ε_t has mean 0 and is serially uncorrelated.

$$\log(A_t) = a_t^* + \phi a_{t-1} + \varepsilon_t \tag{6}$$

Computer RBC

- Representative Household's Objective

$$\begin{aligned} U &= E\sum_{i=0}^{\infty}\beta^i U(C_{t+i}, 1-N_{t+i}), \\ U(C_t, 1-N_t) &= \log(C_t) + \theta\frac{(1-N_t)^{1-\gamma_n}}{1-\gamma_n} \end{aligned}$$

First-Order Conditions

- Define R_t the interest factor received by HH. Firm Behavior implies:

$$R_t = (1-\alpha)\left(\frac{A_t N_t}{K_t}\right)^{\alpha} + (1-\delta) \tag{7}$$

$$W_t = \alpha A_t^{\alpha}\left(\frac{K}{N_t}\right)^{1-\alpha} \tag{8}$$

- Household behavior

$$\frac{1}{C_t} = \beta E_t\left[\frac{R_{t+1}}{C_{t+1}}\right] \tag{9}$$

$$\frac{W_t}{C_t} = \theta(1-N_t)^{-\gamma_n} \tag{10}$$

- From (10), γ_n is inversely related to the elasticity of labor supply.

Summary of the Model

- The numbered equations constitute a system of seven equations in the seven variables $Y_t, K_t, N_t, C_t, R_t, W_t$ and A_t.
- Non-linear equations but that is not such a big problem
- Expectations (forward looking behavior) are the bigger problem - how the economy behaves today depends on how agents think the economy will behave in the future for all possible outcomes:
  - Typically rational expectations
  - Agent-based models in which agents predict according to some of pre-specified rules are easier to solve

Solution

- Methods that focus on the first-order conditions:
  - projection methods
  - perturbation methods
- Methods that are based on the Bellman equation
  - similar to projection methods
  - often slower, but can handle more complex models (e.g. discontinuities)
- How linearisation deals with this? How to think about the problem in terms of policy functions

Solution: Objective of 1st-order perturbation

- The model FOCs are:

$$E_t\left[f\left(g\left(x_t\right)\right)\right]=0 \tag{11}$$

- $g\left(x_t\right)$ is the policy function which we want to know
  - It is a function of the state variables only
- Obtain linear approximations to the policy functions that satisfy the first-order conditions

Solution: The (log-)Linearisation Strategy

- Our goal: Solve for the dynamic path of Y_t, C_t, N_t, etc., for any realization of the series of technology shocks ε_t.
- We do the following:
  1. Solve for the non-stochastic BGP.
  2. Rewrite model as log-deviations from the non-stochastic BGP
  3. Study an alternative model that is: (i) log-linear, and (ii) an approximation of the original model around the non-stochastic BGP.
  4. Interpretation and calculation are made easier, if the equations are linear in percent deviations from the steady state.
- There are two ways to perform the steps above:
  - The easy way - use a computer
  - The hard way - do it by hand

The Non-Stochastic Growth Path

- On the non-stochastic BGP
  - Y^*, K^*, C^*, and W^* all grow at rate $G \equiv A^*_{t+1}/A^*_t$.
  - R^* and N^* are constant.
- From $\frac{1}{C_t} = \beta E_t\left[\frac{R_{t+1}}{C_{t+1}}\right]$,

$$R^* = \frac{G}{\beta}$$

- From $R_t = (1-\alpha)A_t^\alpha\left(\frac{N_t}{K_t}\right)^\alpha + (1-\delta)$, combined with $R = \frac{G}{\beta}$,

$$\frac{K_t^*}{A_t^* N^*} = \left(\frac{1-\alpha}{G/\beta-(1-\delta)}\right)^{1/\alpha}$$

- From $Y_t = K_t^{1-\alpha}(A_t N_t)^\alpha$,

$$\frac{Y_t^*}{A_t^* N^*} = \left(\frac{1-\alpha}{G/\beta-(1-\delta)}\right)^{(1-\alpha)/\alpha}$$

- (Note from the last two we have Y_t^*/K_t^*)

The Non-Stochastic Growth Path

- From $K_{t+1} = (1-\delta)K_t + Y_t - C_t$,

$$\frac{C_t^*}{K_t^*} = 1-\delta-G+\frac{Y_t^*}{K_t^*} \text{ and } \frac{C_t^*}{Y_t^*} = \frac{K_t^*}{Y_t^*}\left(1-\delta-G+\frac{Y_t^*}{K_t^*}\right)$$

- From $W_t = \alpha Y_t/N_t$,

$$\frac{W_t^*}{C_t^*} = \frac{\alpha}{N^*}\frac{Y_t^*}{C_t^*}$$

- From $\frac{W_t}{C_t} = \theta(1-N_t)^{-\gamma_n}$, combined with the previous equation,

$$\frac{\alpha}{N^*}\frac{Y_t^*}{C_t^*} = \theta(1-N_t^*)^{-\gamma_n},$$

  which can now be solved for N^*

Log-Linearization

- Lower-case letters are log-deviations from non-stochastic BGP values. E.g. $y_t = \log\left(\frac{Y_t}{Y_t^*}\right) = \log(Y_t) - \log(Y_t^*)$.
- Log-linearization of $Y_t = K_t^{1-\alpha}(A_t N_t)^\alpha$

$$Y_t^* e^{y_t} = K_t^{*1-\alpha} e^{(1-\alpha)k_t}(A_t^* N_t^*)^\alpha e^{\alpha(a_t+n_t)}$$

  Or

$$y_t = (1-\alpha)k_t + \alpha(a_t+n_t)$$

  (this is exact!).

Log-Linearization

- Log-linearization of $K_{t+1} = (1-\delta)K_t + Y_t - C_t$:

$$K_{t+1}^* e^{k_{t+1}} = (1-\delta)K_t^* e^{k_t} + Y_t^* e^{y_t} - C_t^* e^{c_t}$$

- When x is small, the following approximation works

$$e^x \approx 1 + x.$$

- Hence, a log-linear approximation of the resource constraint is

$$K_{t+1}^* + K_{t+1}^* k_{t+1} = (1-\delta)K_t^* + (1-\delta)K_t^* k_t + Y_t^* + Y_t^* y_t - C_t^* - C_t^* c_t$$

- Since $K_{t+1}^* = (1-\delta)K_t^* + Y_t^* - C_t^*$:

$$K_{t+1}^* k_{t+1} = (1-\delta)K_t^* k_t + Y_t^* y_t - C_t^* c_t$$

  or

$$\frac{K_{t+1}^*}{K_t^*} k_{t+1} = (1-\delta)k_t + \frac{Y_t^*}{K_t^*} y_t - \frac{C_t^*}{K_t^*} c_t$$

- This is linear in k_{t+1}, k_t, y_t, and c_t - the three ratios are constants

Log-Linearization

- Log-linearization of $R_t = (1-\alpha)A_t^{\alpha}\left(\frac{N_t}{K_t}\right)^{\alpha} + (1-\delta)$

$$R^* e^{r_t} = (1-\alpha)\left(\frac{A_t^* N_t^*}{K_t^*}\right)^{\alpha} e^{\alpha(a_t+n_t-k_t)} + (1-\delta),$$

  which is approximately

$$R^* r_t = (1-\alpha)\alpha\left(\frac{A_t^* N_t^*}{K_t^*}\right)^{\alpha}(a_t+n_t-k_t).$$

Log-Linearization

- Log-linearization of $W_t = \alpha A_t^{\alpha}\left(\frac{K}{N_t}\right)^{1-\alpha}$

$$w_t = \alpha a_t + (1-\alpha)(k_t - n_t)$$

  (exact).

Log-Linearization

- Log-linearization of $\frac{W_t}{C_t} = \theta(1-N_t)^{-\gamma_n}$:

$$\begin{aligned}\frac{W_t^* e^{w_t}}{C_t^* e^{c_t}} &= \theta(1-N^* e^{n_t})^{-\gamma_n}\\ \frac{W_t^* + W_t^* w_t}{C_t^* + C_t^* c_t} &= \theta(1-N^* - N^* n_t)^{-\gamma_n}\end{aligned}$$

- Problem: $(1-N^* e^{n_t})$ does not conveniently break down into $(1-N^*)(1+$something linear in $n_t)$
- Take a first order Taylor expansion around $w_t = c_t = n_t = 0$:

$$\begin{aligned}\frac{W_t^*}{C_t^*} + \frac{W_t^*}{C_t^*}w_t - \frac{W_t^*}{C_t^*}c_t &= \theta(1-N^*)^{-\gamma_n} + \theta\gamma_n(1-N^*)^{-\gamma_n-1}N^* n_t\\ \frac{W_t^*}{C_t^*}(w_t - c_t) &= \theta\gamma_n(1-N^*)^{-\gamma_n}\frac{N^*}{1-N^*}n_t\end{aligned}$$

$$\Rightarrow w_t - c_t = \gamma_n \frac{N^*}{1-N^*} n_t$$

Log-Linearization

- Log-linearization of $\frac{1}{C_t} = \beta E_t\left[\frac{R_{t+1}}{C_{t+1}}\right]$

$$\frac{1}{C_t^* e^{c_t}} = \beta E_t\left[\frac{R^* e^{r_{t+1}}}{C_{t+1}^* e^{c_{t+1}}}\right]$$

  which is exactly

$$\frac{1}{e^{c_t}} = E_t\left[\frac{e^{r_{t+1}}}{e^{c_{t+1}}}\right]$$

  and which, with the usual approximation, we can rewrite

$$\frac{1}{1+c_t} = E_t\left[\frac{1+r_{t+1}}{1+c_{t+1}}\right].$$

  Once again, however, this is not yet nice and linear. Let's do first-order Taylor around $c_t = r_{t+1} = c_{t+1} = 0$. We get:

$$-c_t = E_t\left(r_{t+1} - c_{t+1}\right)$$

Log-Linearization Recipe

Recipe to log-linearize some relation $g(X_t) = 0$ with $g(X^*) = 0$

1. Rewrite as $g(X^* e^{x_t}) = 0$, where $x_t = \log(X_t) - \log(X^*)$
2. Can you use $g(X^*) = 0$ to get a linear equation in x_t?
    - If yes, stop. e.g. worked with $Y_t = K_t^{1-\alpha}(A_t N_t)^\alpha$ and $W_t = \alpha A_t^\alpha \left(\frac{K}{N_t}\right)^{1-\alpha}$. The resulting expression is exact.
    - If not, continue.
3. Approximate everywhere e^{x_t} with $1 + x_t$.
4. Again try to use $g(X^*) = 0$ to make this linear in x_t. Did it work?
    - If yes, stop. This worked with $K_{t+1} = (1-\delta)K_t + Y_t - C_t$ and with $R_t = (1-\alpha)A_t^\alpha \left(\frac{N_t}{K_t}\right)^\alpha + (1-\delta)$.
    - If not, continue.
5. Take a first-order Taylor expansion around $x_t = 0$ and use (as always) $g(X^*) = 0$. This always works.

Summary of the Log-Linear Model

Production Function	$y_t = (1-\alpha)k_t + \alpha(a_t + n_t)$
Resource Constraint	$\frac{K_{t+1}^*}{K_t^*} k_{t+1} = (1-\delta)k_t + \frac{Y_t^*}{K_t^*} y_t - \frac{C_t^*}{K_t^*} c_t$
Interest Rate	$R^* r_t = (1-\alpha)\alpha \left(\frac{A_t^* N_t^*}{K_t^*}\right)^\alpha (a_t + n_t - k_t)$
Wage	$w_t = \alpha a_t + (1-\alpha)(k_t - n_t)$
Labor-Leisure Choice	$w_t - c_t = \gamma_n \frac{N^*}{1-N^*} n_t$
Euler Equation	$-c_t = E_t\left(r_{t+1} - c_{t+1}\right)$
Technology	$a_t = \phi a_{t-1} + \varepsilon_t$

- "$*$" variables $\equiv$ parameters
- 7 first-order linear stochastic difference equations in the 7
- $y_t, k_t, n_t, c_t, r_t, w_t$ and a_t

Solving the Log-Linearized Model

- The same that we always want - IRF: solve for the (path of the) endogenous variables as functions of the model's parameters $(\alpha, \beta, \gamma_n, \delta, \phi, G, \theta)$ and the (path of the) exogenous shock process ε_t
- Believe it or not someone has written open-source software that will do ALL OF THIS for you. You can get it for free at `www.cepremap.cnrs.fr/dynare/`
- Actually dynare will do much more for you than what we did
    - It will solve for higher-order (not just first order) approximation
    - It will estimate the parameters of the model from time series data (instead of relying on a calibration).
    - (Indeed estimation of DSGE is increasingly popular and may end up replacing calibration)

Calibration

- The time series properties of the model depend on the ηs
- which in turn depend on the models parameters $(\alpha, \beta, \gamma_n, \delta, \phi, G, \theta)$.
- Calibration is about picking numbers for these parameters.
- There are many different ways of doing this.

Calibration

For G:

- The average annual growth rate of GDP in the US has been roughly 2% historically.
- If interpreted as BGP growth rate, annual growth rate of $A^* = 1.02$
- Make this a quarterly growth rate:

$$G = 1.005$$

Calibration

For β

- The real interest rate in the US (if such a thing exists) has averaged roughly 6%
- We take this as an estimate of the real interest rate on the non-stochastic BGP... quarterly rate for model is $R^* = 1.015$
- In steady state we have $R^* = G/\beta$, this gives a choice for β.
- However, once you have R^*, β does not appear anywhere else in the evaluation of steady state quantities, or in the log-linearized model.
- So we can just use $R^* = 1.015$ throughout.

Calibration

For θ

- An introspective estimate (not mine!) suggests that households allocate about 1/3 of their time endowment to market activities.
- This means $N^* = 1/3$.
- Recall that N^* is pinned down by the equation:

$$\frac{\alpha}{N^*}\frac{Y_t^*}{C_t^*} = \theta(1 - N_t^*)^{-\gamma_n}$$

  so plugging in $N^* = 1/3$, as well as the value of $\frac{Y_t^*}{C_t^*}$ implied by the previous choices of parameters implicitly delivers our choice of θ.
- Again though, this is the only place we see θ so we just use $N^* = 1/3$ throughout.

Calibration

For α:

- The labor share has been roughly constant at 2/3 over long-run US history. Hence $\alpha = 0.667$.

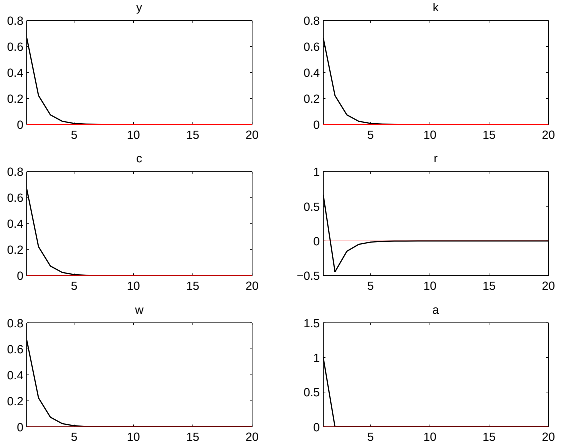
For δ:

- Based on NIPA statistics, a consensus view is that the depreciation rates is roughly 10% at an annual rate, which implies $\delta = 0.025$.

For γ_n and ϕ:

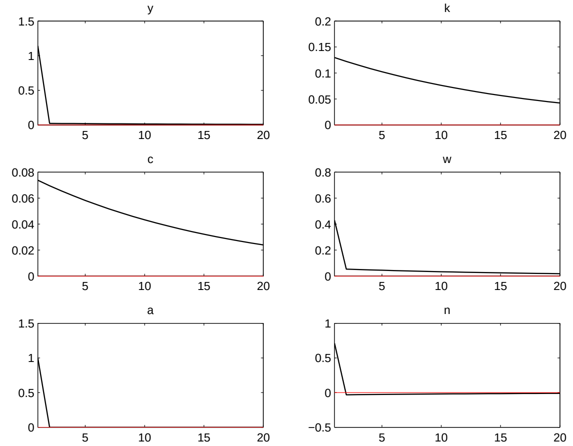
- Campbell does not calibrate these parameters but experiments with different values.

$\delta = 1, \gamma = 1, \phi = 0$



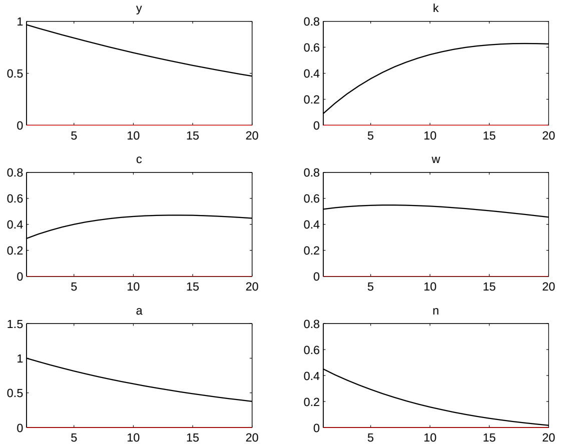
This is what we are trying to improve upon.

Incomplete depreciation: $\delta = 0.025, \gamma = 1, \phi = 0$



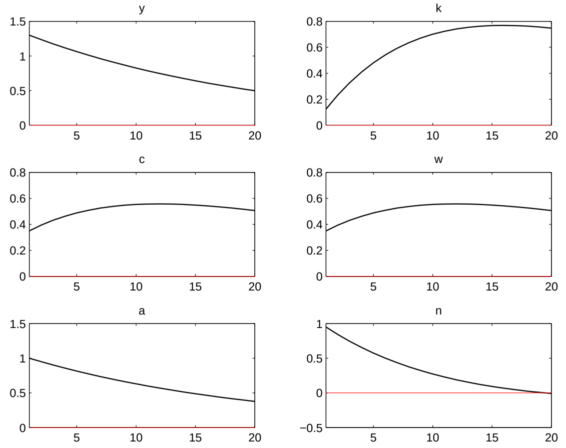
- Employment now responds but not enough (Insufficient amplification)
- Business cycle too short. (Insufficient persistence)
- Consumption response too small
- Wage may be too procyclical

Persistent Shock: $\delta = 0.025, \gamma = 1, \phi = 0.95$



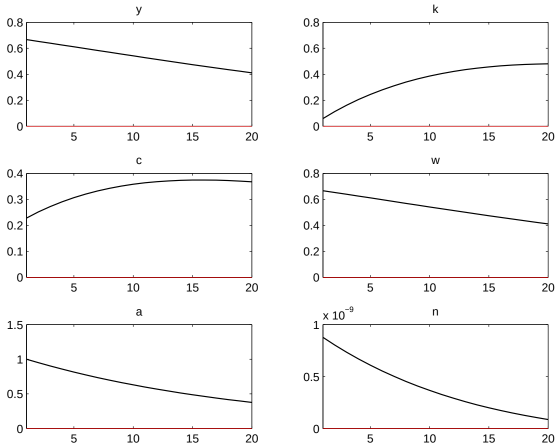
- Still not much amplification: Output inherits the shock's persistence
- Consumption dynamics may be a bit unrealistic (and small contemporaneous comovement)
- Wage even more procyclical

Very Elastic Labor Supply: $\delta = 0.025, \gamma = 0, \phi = 0.95$



- Desired labor supply response
- Nothing much else changes

Very Inelastic Labor Supply: $\delta = 0.025, \gamma = 1000000000, \phi = 0.95$



- No employment response
- Huge wage increase

Transparency and Deliberation

Mario Draghi (2013): "It would be wise to have a richer communication about the rationale behind the decisions that the governing council takes."

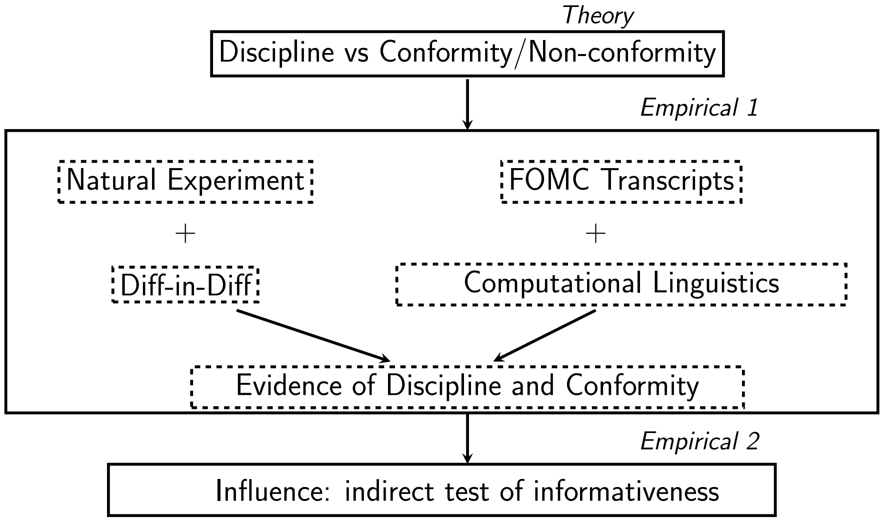
	Fed (2014)	BoE (2014)	ECB (2014)
Minutes?	✓	✓	X
Transcripts?	✓	X	X

April 30, 2014: BoE to review of non-release of transcripts
July 3, 2014: ECB to release account of meetings

Specific goal of this Paper

We want to study how transparency affects FOMC deliberation.
⇒ how is *internal deliberation* affected by greater *external communication*?

The outline of our paper



Theory

Discipline vs Conformity/Non-conformity

Empirical 1

Natural Experiment + Diff-in-Diff

FOMC Transcripts + Computational Linguistics

Evidence of Discipline and Conformity

Empirical 2

Influence: indirect test of informativeness

Our Natural Experiment

- FOMC meetings were recorded and transcribed from at least the mid-1970's in order to assist with the preparation of the minutes.
- This fact was unknown to committee members prior to November 1993
- October 1993: Alan Greenspan acknowledged the transcripts' existence to the Senate Banking Committee.
- The Fed then quickly agreed:
  - To begin publishing them with a five-year lag.
  - To publish the back data
- A natural experiment we exploit to assess the effect of transparency
  - Prior to Nov 1993: Discussion took place under the assumption that individual statements would not be on the public record
  - After Nov 1993: Each policy maker knew that every spoken word would be public within five years.

Summary table of effects of transparency

Discipline	Conformity
↑ economics topic coverage in FOMC1	↓ statements in FOMC2
↑ references to data topics in FOMC1	↓ questions in FOMC2
↑ use of numbers in FOMC1	↓ distance from Greenspan in FOMC2
	↓ economics topic coverage in FOMC2
↑ economics topic percentage in FOMC2	

- Discipline tends to increase informativeness, conformity to decrease it. Which effect dominates?
- If member's statement more informative, it should drive the debate more.

Text as Data

We can represent text as a sequence of words and punctuation.

'Today' 'I' 'am' 'in' 'a' 'workshop' ',' 'and' 'having' 'fun' '.' '¶' 'Aren't' 'you' '?'

This sort of data is unstructured in the sense that the information we want is not readily accessible.

Example: does the document contain the word 'workshop'? Need to transform each document into a 0 or a 1.

Broadly speaking, text mining is the study of the quantitative representation of text.

Relevant Information

Any text mining algorithm will throw away some information.

Deciding which information is important and which is extraneous depends on context. This is the art of data science.

For example, the *bag of words* representation removes all punctuation, and treats every word as independent. Sufficient statistic for text is histogram counts of unique words in data.

The above sentence is then equivalent to

'Today' 'I' 'am' 'in' 'a' 'workshop' 'and' 'you' 'Aren't' 'having' 'fun'

Clearly some important meaning is gone. Does this matter?

Two Situations

Suppose you want to predict how many classes Warwick runs per year on data science with blog posts. The bag of words representation is useful.

Suppose you want to predict which word is most likely to complete the blank:

'Today' 'I' 'am' 'in' 'a' 'workshop' ',' 'and' 'having' 'fun' '.' '¶' 'Aren't' '__' '?'

The bag of words is useless. Representing text as a sequence of word pairs (bigrams) would be better. For example

('Today' 'I') ('I' 'am') ('am' 'in') ('in' 'a') ('a' 'workshop') ('workshop' 'and') ('and' 'having') ('having' 'fun') . . .

Note we still lose information.

An Ideal Dataset

The perfect dataset for analysis might look like the following

Speaker	Date	Location	Text
A	Jan 14	LSE	...
B	Jun 14	LSE	...
C	Jun 14	UCL	...
B	Aug 14	Oxford	...
...	...	...	...

The text is unstructured data, but it lives a structured, rectangular database with associated metadata.

Bad news: Getting to this point can be quite tedious. Data science is 80% data cleaning, 20% analysis.

Possible Data Sources

We have used data from the following sources:

1. Static HTML web pages
2. Word documents
3. PDF documents ⟶ plain text files via OCR
4. Web applications' APIs

Key problems:

1. Separate metadata from text
2. Break texts up at appropriate points (paragraphs, bullet points, etc)
3. Remove graphs and charts

Unfortunately, most authors write documents in highly specific ways, requiring bespoke solutions.

Option 1

HTML pages and Word documents carry within them useful tags.

Modules exist to break up documents according to these tags, and if you're lucky these tags will correspond to the metadata you want.

For example, Python modules *beautiful soup* and *docx*.

Option 2

Decide on an organizational standard for document construction.

For example, Facebook and Twitter have standard fields into which all text must go.

Analogy at Bank of England would be to have common template across divisions for writing policy notes.

At the moment, Filesite contains a wealth of information, but the formats are quite heterogeneous.

Option 3

Ultimately one may have no option other than manual extraction.

Cheap and reliable services exist for such work in developing countries.

Text Processing

Suppose we have a clean database in front of us.

Before we perform any mathematical operations on text, we have to process the strings.

Basic steps:

1. Break statements into tokens
2. Decide whether and how you want to transform tokens
3. Decide which tokens you want to keep

Many of these operations are implemented in Python’s Natural Language Toolkit

Tokenization

The most basic way of forming tokens is to split on whitespace and punctuation.

‘I think, therefore I am.’ ⟶ ‘I’ ‘think’ ‘,’ ‘therefore’ ‘I’ ‘am’ ‘.’

One issue is that sometimes punctuation joins words together, or there are natural phrases.

We probably don’t want to represent:

- ‘aren’t’ as ‘aren’ ‘’’ ‘t’
- ‘risk-weighted assets’ as ‘risk’ ‘-’ ‘weighted’ ‘assets’
- ‘fed funds rate’ as ‘fed’ ‘funds’ ‘rate’

Options:

1. Remove all punctuation that has an alphabetic character on either side before tokenizing
2. Make custom list of transformations
3. Ignore the problem

Token Transformation: Case Folding

A common transformation is to convert all tokens to lowercase:

‘I’ ‘think’ ‘,’ ‘therefore’ ‘I’ ‘am’ ‘.’ ⟶ ‘i’ ‘think’ ‘,’ ‘therefore’ ‘i’ ‘am’ ‘.’

Main issue is that capitalization sometimes affects meaning, for example ‘US’ $\neq$ ‘us’ nor ‘CAT’ $\neq$ ‘cat’.

But even Google can’t fully overcome this problem! (Try a search for ‘C.A.T.’).

Token Transformation: Linguistic Roots

Option #1: Lemmatizing (first-best). Requires part-of-speech tagging first.

Converts ‘ran’ to ‘run’ if tagged as verb.

Option #2: Stemming (dirty but fast). Deterministic algorithm to remove ends from words. ‘prefer’, ‘prefers’, and ‘preference’ all become ‘prefer’.

Not necessarily English word. ‘inflation’ becomes ‘inflat’.

Choosing Words to Keep

A large percentage of words in any text are articles and prepositions like ‘a’, ‘the’, ‘to’, etc.

Stopwords are common words that we remove from text as part of pre-processing. Standard lists are available.

Also common to drop words that appear in few and many documents.

Fancier option is to use tf-idf weights.

Also common to remove punctuation, but these can also be informative depending on context.

Bag of Words and Dimensionality

After all of these steps, we have a database of lists of terms and are ready for analysis.

The bag of words representation is a histogram with dimensionality equal to the vocabulary size:

1. Usually thousands of dimensions of variation
2. Sparsity: most documents in corpus have 0 count for most terms.

How can we address the dimensionality problem. Two options:

1. Dictionary methods
2. Latent variable methods

Dictionary Methods

Analysis by dictionary methods has two steps.

First, define a list of key words that capture content of interest.

Second, represent each document in terms of the (normalized) frequency of words in the dictionary.

For example, let the dictionary be $D = \{\text{labor}, \text{wage}, \text{employ}\}$.

One could then represent each document i as

$$d_i = \frac{\#\ \text{labor occurrences} + \#\ \text{wage occurrences} + \#\ \text{employ occurrences}}{\text{total words in document}}$$

Example

Example: Tetlock, et al. (2008). "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *Journal of Finance*.

Example: Tucket (2013). "Irreducible Uncertainty and its Implications: A Narrative Action Theory for Economics."

Limitations of word counting?

Inverse Document Frequency

- Words that appear frequently are not informative—"inflation"
- Rare words are valuable for discriminating in terms of content—"deflation"
- Let df_v be the number of documents in which word v appears
- The *inverse document frequency* is $\text{idf}_v = \log\left(\frac{N}{df_v}\right)$
- Each word $v \in 1, \ldots, V$ is assigned a score
- Motivation for log: Zipf's Law

Example for 1,000,000 Documents

term	df_v	idf_v
calpurnia	1	6
animal	100	4
sunday	1000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

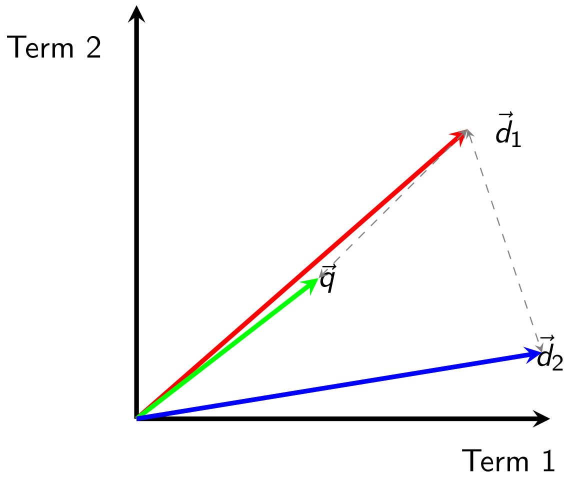
TF-IDF weighting

- Idea of idf is not that collection frequencies don't matter, just that they should be scaled up or down with an appropriate weight
- Let f_v be the collection frequency of term v. Define the collection level *term frequency-inverse document frequency* as $\text{tf-idf}_v = \log(1 + f_v) \times \text{idf}_v$.

Vector Space Model

- One can also use tf-idf to represent each document $i \in 1, \ldots, N$ as a vector $\vec{d}_i$ of dimension V.
- Let the vth element of this vector be $\text{tf-idf}_{iv} = \log(1 + f_{iv}) \times \text{idf}_v$.
- A straightforward extension of word counting for economics research is to base analysis on how the vth term varies across across i.
- How to compare documents across all dimensions?

Why (Euclidean) Distance is a Bad Idea

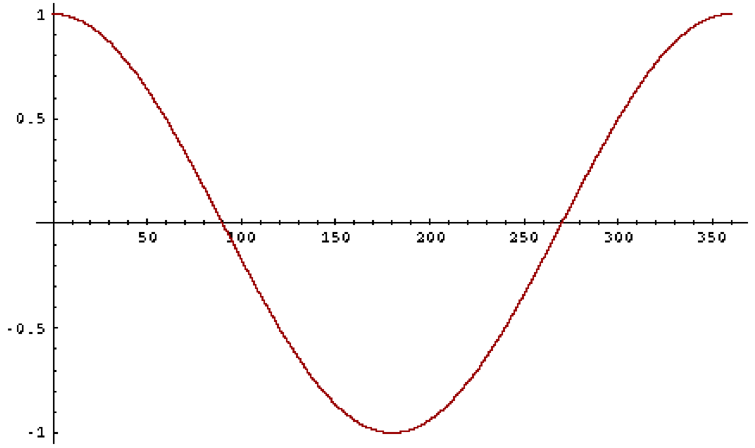


The Euclidean distance of $\vec{q}$ and $\vec{d}_1$ is large although the distribution of terms in q and d_2 are very similar.

Angle versus Distance

- Thought experiment: take a document d and append it to itself. Call this document d'. d' is twice as long as d.
- "Semantically" d and d' have the same content.
- The angle between the two documents is 0.
- But Euclidean distance between the two documents can be quite large.

Cosine



Cosine Similarity

- Define the cosine similarity between documents i and j as

$$CS(i,j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\left\|\vec{d}_i\right\| \left\|\vec{d}_j\right\|}$$

- $CS(i,j) \in [0,1]$
- Very popular in information retrieval
- See `http://alex2.umd.edu/industrydata/` for economics application

Topic Models

Latent variable models view the histogram of words (with dimensionality V) as coming from a set of topics with much lower dimensionality ($K << V$).

General idea is that related words should map back into single theme

$$\{\text{labor}, \text{wage}, \text{employ}\} \rightarrow \{\text{labor markets}\}$$

A topic β_k is a probability distribution over the V terms in the data.

Flexible structure is important. For example, topics about animals and war. How to assign the word “tank”?

The Latent Dirichlet Allocation (LDA) model

- Blei, Ng and Jordan (2003) cited 10,000+ times: new to economics.
- LDA (and its extensions) estimates what fraction of each document in a collection is devoted to each of several “topics.”
  - JSTOR example
- Great promise for economics more broadly.
- LDA is an unsupervised learning approach - we don’t set probabilities

1. Start with words in statements
2. Tell the model how many topics there should be
   - Perplexity scores
3. Model will generate β_K **topic distributions**
   - the distribution over words for each topic
4. Model also generates θ_d **document distributions**

Example statement: Yellen, March 2006, #51

Raw Data → Remove Stop Words → Stemming → Multi-word tokens = Bag of Words

We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.

Example statement: Yellen, March 2006, #51

Raw Data → Remove Stop Words → Stemming → Multi-word tokens = Bag of Words

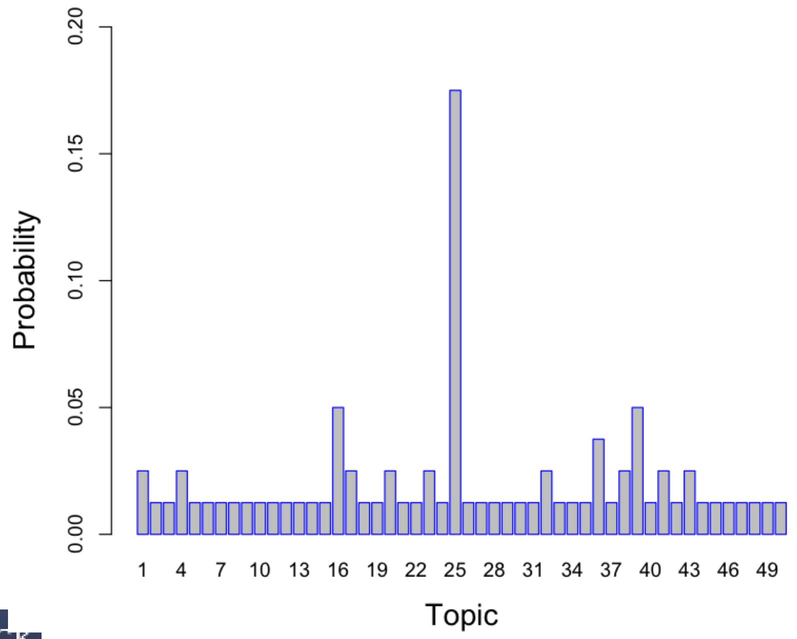
noticed change relationship between core CPI
chained core CPI suggested maybe something going
relating substitution bias upper level index focused
nonmarket component PCE wondered something
unusual happening core CPI relative measures

Example statement: Yellen, March 2006, #51

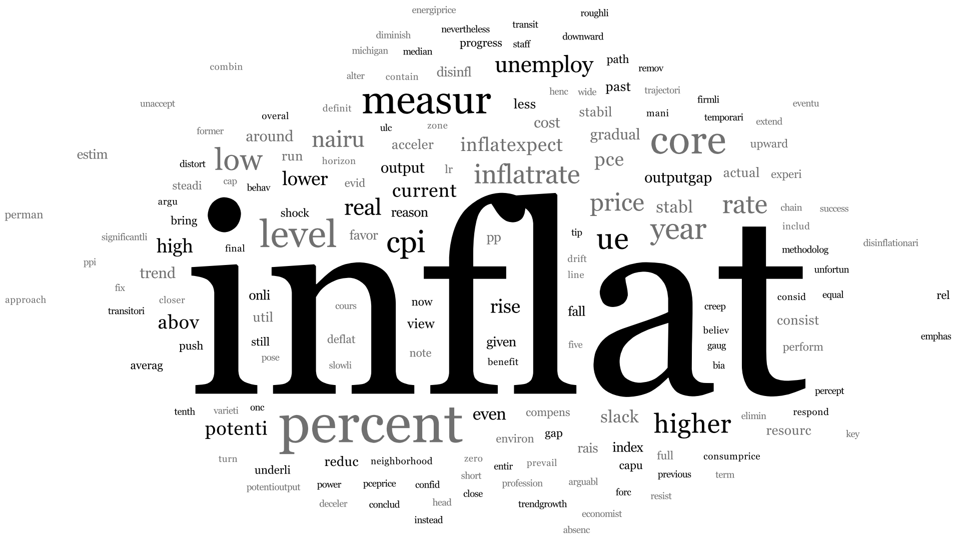
Raw Data → Remove Stop Words → Stemming → Multi-word tokens = Bag of Words

notic chang relationship between core CPI
chain core CPI suggest mayb someth go
relat substitut bia upper level index focus
nonmarket compon PCE wonder someth
unusu happen core CPI rel measur

Predictive Posterior Distribution



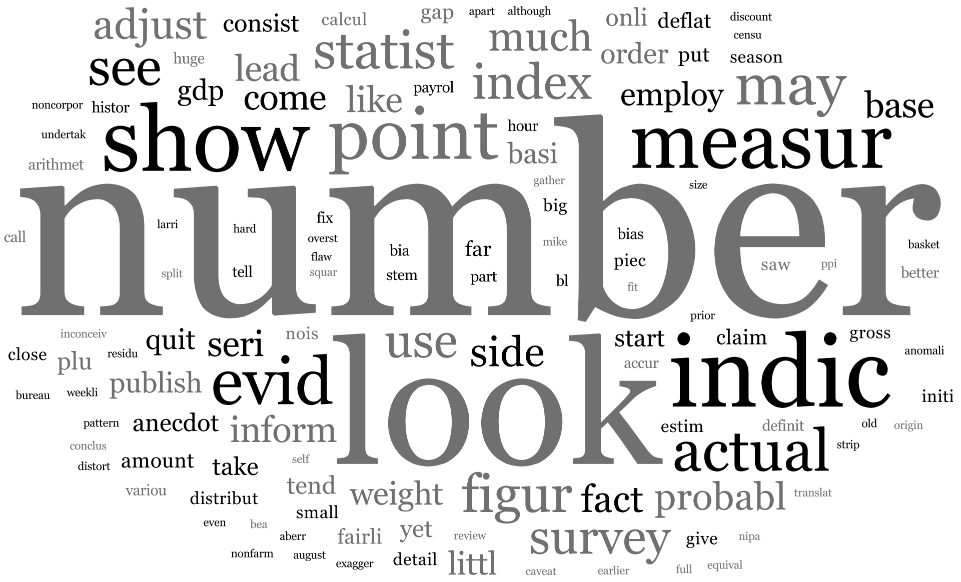
Topic 25



Advantage of Flexibility

‘measur’ has probability 0.026 in topic 25, and probability 0.021 in topic 11

Topic 11



Advantage of Flexibility

‘measur’ has probability 0.026 in topic 25, and probability 0.021 in topic 11.

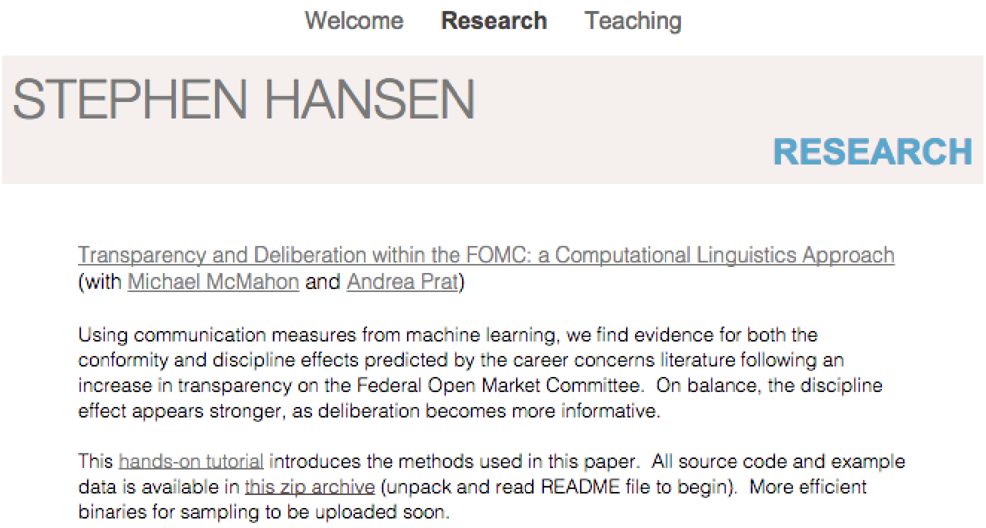
It gets assigned to 25 in this statement consistently due to the presence of other topic 25 words.

In statements containing words on evidence and numbers, it consistently gets assigned to 11.

Flexibility important!

A shameless plug

`http://www.econ.upf.edu/~shansen/Research.html`



STEPHEN HANSEN

RESEARCH

Transparency and Deliberation within the FOMC: a Computational Linguistics Approach
(with Michael McMahon and Andrea Prat)

Using communication measures from machine learning, we find evidence for both the conformity and discipline effects predicted by the career concerns literature following an increase in transparency on the Federal Open Market Committee. On balance, the discipline effect appears stronger, as deliberation becomes more informative.

This hands-on tutorial introduces the methods used in this paper. All source code and example data is available in this zip archive (unpack and read README file to begin). More efficient binaries for sampling to be uploaded soon.

To Cover

1. Matlab basics
   1.1 The screens
   1.2 *m*-files
   1.3 functions
2. Basic Matrices
   - Inputting a matrix
   - Useful matrices
   - Matrix coordinates
3. Plots in Matlab - especially 3-D

Matlab I

The screens consist of a number of windows (like Stata):

- Command Window
- History Window
- Workspace / current directory
- Editor
- Help
- Graphics window

Matlab Editor

The editor is like any standard text editor, but integrates with MATLAB.

- F5 - Saves and runs your current m-file.
- F9 - Evaluates a selection of the m-file and ignores those parts that were not selected. Can be used for a single line of text.
- ... - For long lines of code, use an ellipsis to tell MATLAB to continue on the next line. Anything after the elipsis is regarded as a comment.
- TAB - Use the TAB key to indent your code. Code that exists between an if or for and the corresponding end should be indented. This will help both you and others understand your code. (as with Stata last week - it is only for ease)
- ; - indicates “no output” - important when the matrix is large.

Matlab Functions

- There are a number of in-built functions, but similarly you can create your own very easily;
- Save the m-file as the function name
- EXAMPLE: welfun.m
  - inputs = c, time
  - output = $\sum_{j=1}^{time} j.c^j$
- But you can actually write more useful ones!

Inputting a Matrix

Matlab carries out most commands using matrices.
If I want to input a matrix I simply use:

- square brackets
- , to indicate different inputs (space usually works)
- ; to indicate the end of a row.

A = [1, 2, 3; 1, 2, 3; 1, 2, 2];

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 2 \end{bmatrix}$$

Inputting a Matrix

Then usual commands work well:

- transpose $\implies$ A'
- inverse $\implies$ $inv(A)$
- matrix multiplication $\implies$ $A * B$
- And: $+, -, /$
- With square matrices: ^

Useful Matrices

Where X is a number, or a set of coordinates:

- identity matrix. $\implies$ eye(X)
- matrix of zeros. $\implies$ zeros(X)
- matrix of ones. $\implies$ ones(X)
- Uniform random variable. $\implies$ rand(X)
- randn(X). $\implies$ Normal random variable.

Matrix Manipulation

- If we have a matrix, we can refer to individual rows and columns:
- $A(i, j)$ will tell us the entry at row i and column j of matrix A
- ...using a : refers to all rows $A(:, j)$ or columns $A(i, :)$
- $size(A)$ $\implies$ Returns a row-vector giving the size of each dimension of A.
- $length(A)$ $\implies$ Returns the number of columns of A. Or, if A is a column-vector, its height.

Flow Control

- As with Stata programming, you can use:
  - if statements
  - for statements

```
if isequal(A, eye(2));
disp(’A is the 2x2 identity’);
end

for a = 1:3       % a runs from 1 to 3
disp(a^2);       % Display a squared
end;
```

Graphs in Matlab

There are some nice plots that are possible with Matlab - especially its 3-D functionality.

See examples...

END

Questions?