Big Data Analysis of the Macro Economy

Serena Ng
Columbia University

June 2016

Macroeconomic Data

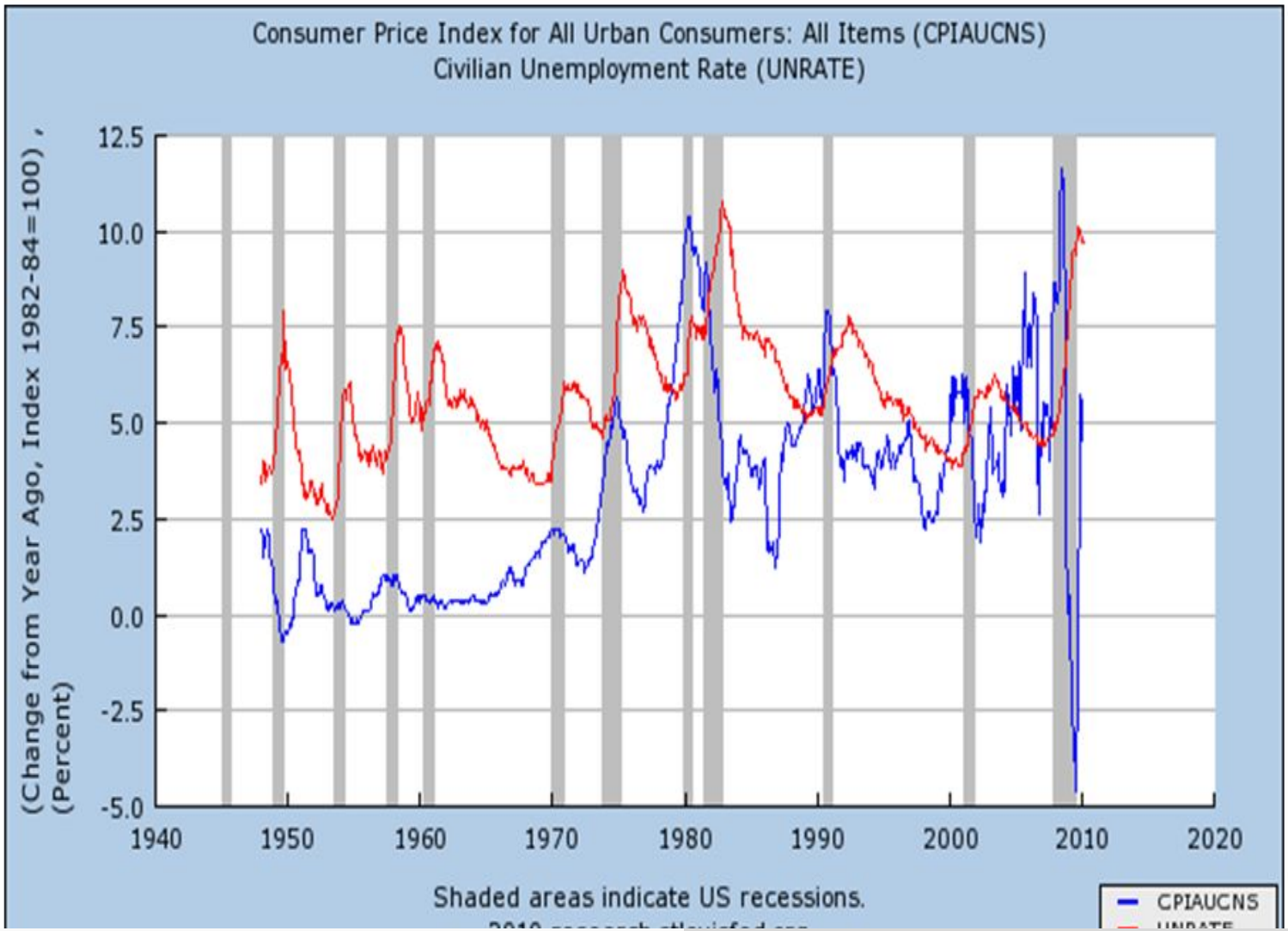
Economic time series: $y_{ti} = \underbrace{T_{ti}}_{\text{trend}} + \underbrace{C_{ti}}_{\text{cycle}} + \underbrace{S_{ti}}_{\text{seasonal}} + \underbrace{I_{ti}}_{\text{irregular}}$.

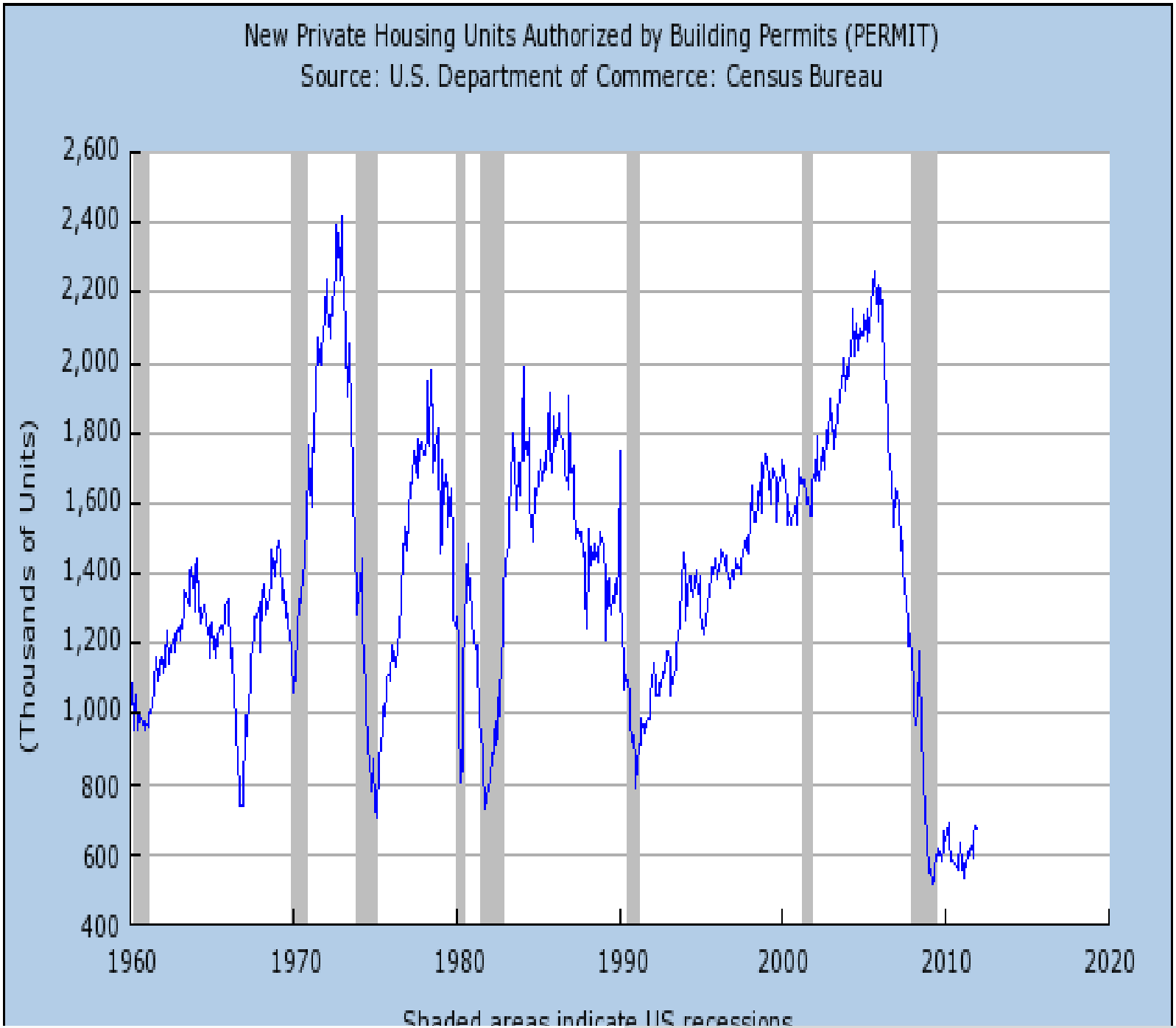
- Macroeconomics: studies causes and consequences of business cycles, ie. ups and downs in economic activity
  - Need to remove T_{ti}, S_{ti}, I_{ti};
  - No unique measure of economic activity.
- Some facts about business cycles:
  - periodicity between 6 and 32 quarters.
  - fluctuations have non-uniform periodicity or amplitude.
  - strong comovements, pervasive.
  - downturns not in a particular industry, or region.

Contractions in the U.S. economy (1900-2011)

Cycle N.	Started	Finished	Total Months	Date S&P Turned Down	Months Before Contraction
1	Sep-1902	Sep-1904	24	Sep-1902	0
2	May-1907	Jun-1908	13	Sep-1906	8
3	Jan-1910	Jan-1912	24	Dec-1909	1
4	Jan-1913	Dec-1914	23	Sep-1912	4
5	Aug-1918	Mar-1919	7	Nov-1916	21
6	Jan-1920	Jul-1921	18	Jul-1919	6
7	May-1923	Jul-1924	14	Mar-1923	2
8	Oct-1926	Nov-1927	13	Feb-1926	8
9	Aug-1929	Mar-1933	43	Aug-1929	0
10	May-1937	Jun-1938	13	Feb-1937	3
11	Feb-1945	Oct-1945	8	Feb-1945	0
12	Nov-1948	Oct-1949	11	May-1946	30
13	Jul-1953	May-1954	10	Dec-1952	7
14	Aug-1957	Apr-1958	8	Jul-1956	13
15	Apr-1960	Feb-1961	10	Jul-1959	9
16	Dec-1969	Nov-1970	11	Nov-1968	13
17	Nov-1973	Mar-1975	16	Dec-1972	11
18	Jan-1980	Jul-1980	6	Jan-1980	0
19	Jul-1981	Nov-1982	16	Nov-1980	8
20	Jul-1990	Mar-1991	8	May-1990	2
21	Mar-2001	Nov-2001	8	Aug-2000	7
22	Dec-2007	Mar-2009	15	Dec-2007	0
23	?	?	?	?	?
		Average	14.5	Average	7.0

Inflation and Unemployment in U.S. 1950-2009




New Private Housing Units Authorized by Building Permits (PERMIT)
Source: U.S. Department of Commerce: Census Bureau
(Thousands of Units)
2,600
2,400
2,200
2,000
1,800
1,600
1,400
1,200
1,000
800
600
400
1960
1970
1980
1990
2000
2010
2020
Shaded areas indicate US recessions

Policy makers and business want to know

- the 'state' of the economy C_t (nowcast)
- where we are in the cycle, where we are going (forecast, anticipate turning points).

Past work: represent cycle by one series: GDP$= \sum_{i=1}^{N} y_{ti}$.
Big data approach: studies C_t common to many y_{ti}.

Big data

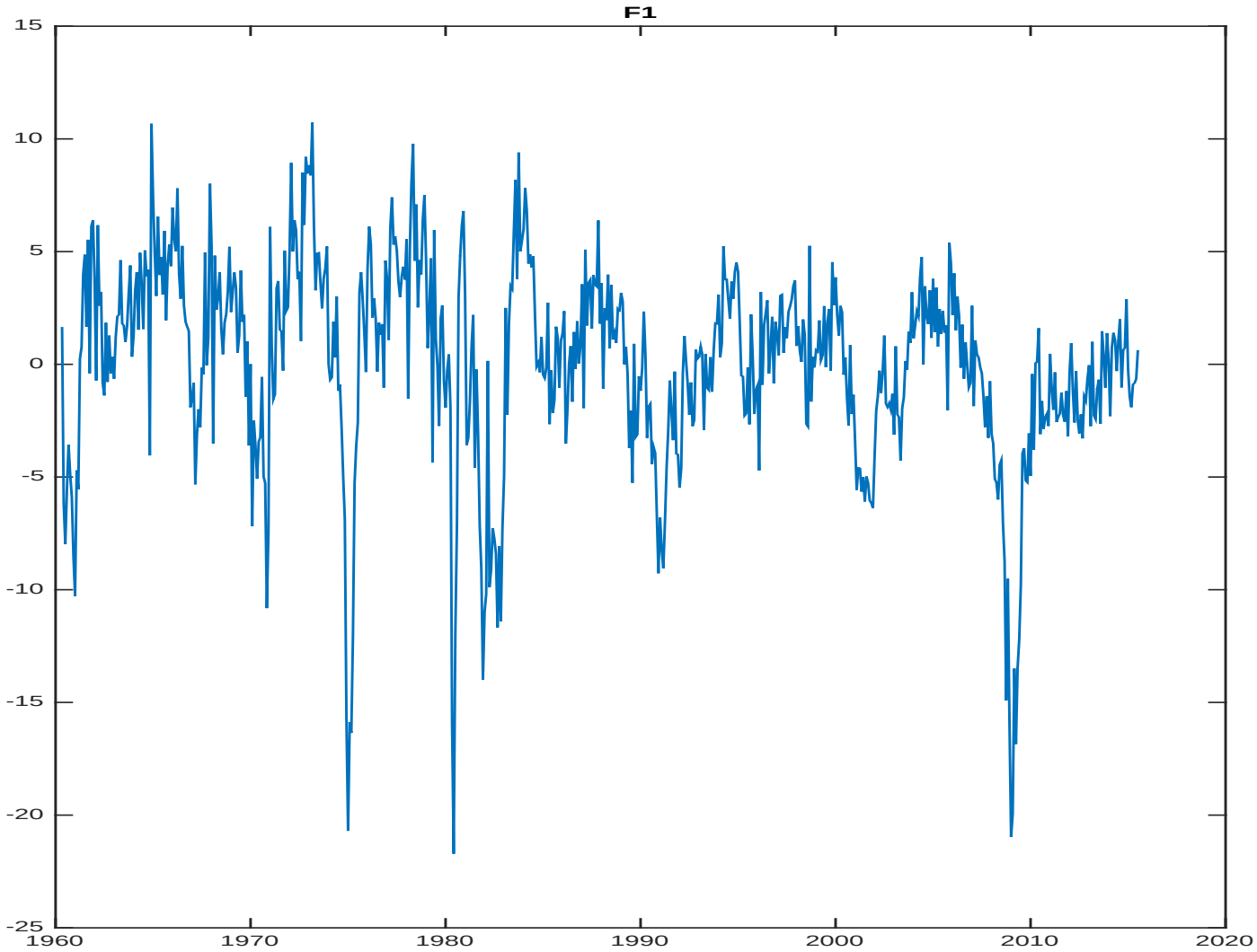
Use common factors to measure economic activity.

- Common variations = low rank component.
- New asymptotic thoery: large T *and* N.
  - $N = 134$ series over $T = 564$ months (1959:1-2015:12).
  - Factor model: $X_{it} = \lambda_i' F_t + e_{it}$

$$\min(N, T)\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{F}_{t,r} - HF_t\right\|^2\right) = O_p(1).$$

  - stimate factors by PCA as guide to C_t
  - $\widehat{F}_1$: loads heavily on production, employment.
  - $\widehat{F}_2$: loads heavily on price variables.

$\widehat{F}_1$ estimated by PCA



PCA lacks interpretation.
CUR isolate important variables.

- 30 variables with top leverage scores:
  - Picked all the usual suspects!
  - NAPM variables, housing starts, term spreads.
  - Very encouraging.
  - But.. picked variables that are too similar. (e.g.one-year and two year interest rates)

Conventional Data

- Drawbacks
  - p and q are indexes, not actually transactions data.
  - not timely, available only months after
  - aggregate time series, loose cross-section information.
  - Non-stationary (unit roots).
- Very convenient
  - The government 'cleans' the data first.
  - Only need to worry about C_t.

Nielsen Scanner Data: 3.6TB

- 3M unique UPCs, 1K+ products, 55 MSA, 35K stores.
- $T = 260$ weeks (2006-2010), span Great Recession.
- Unique features:
  - variables: p_{ti} and q_{ti}: instead of p_{ti} and $v_{ti} = p_{ti}q_{ti}$.
  - frequency: weekly, no time aggregation.
  - spatial: many locations, products.

Goal: extract C_t, the common cyclical variations in this data.

- Aggregate across goods or location gives matrices $A = P$ (price), Q (quantity), V (value).
  - $T \times N$, rows index weeks, columns index store-upc.
- Goal: Construct C_t^a, for $a = p, q, v$ from A, P, V.
- Since C_t^a is common to units in A, I should not need to analyze all the data. But how to subsample?
  - Random projections
  - leverage-score

Issue 1: algorithmic vs. statistical results

- We know, eg. $\|A - P_C A\|_\xi \leq k\sqrt{\log k}\,\|A - P_{U_k} A\|_\xi$
  - Let A_k be $T \times k$: $A_k = U_k D_k V_k^T$.
  - $\widehat{F}^k_{1:r} = (\widehat{F}^k_{1,1:r}, \ldots, \widehat{F}^k_{T,1:r})^T$: first r components from A_k.
  - A desired error rate: $\left\|\widehat{F}^k_{t,1:r} - \widehat{F}^N_{t,1:r}\right\|$ as a function of N, T so that combined with

$$\min(N, T)\left\|\widehat{F}^N_{t,1:r} - HF^N_{t,1:r}\right\|^2 = O_p(1)$$

  some statistical statements about $\widehat{F}^k_{t,1:r}$ can be made.

  Take $C_t = \widehat{F}^k_{1,1}$.

Issue 2: data not iid

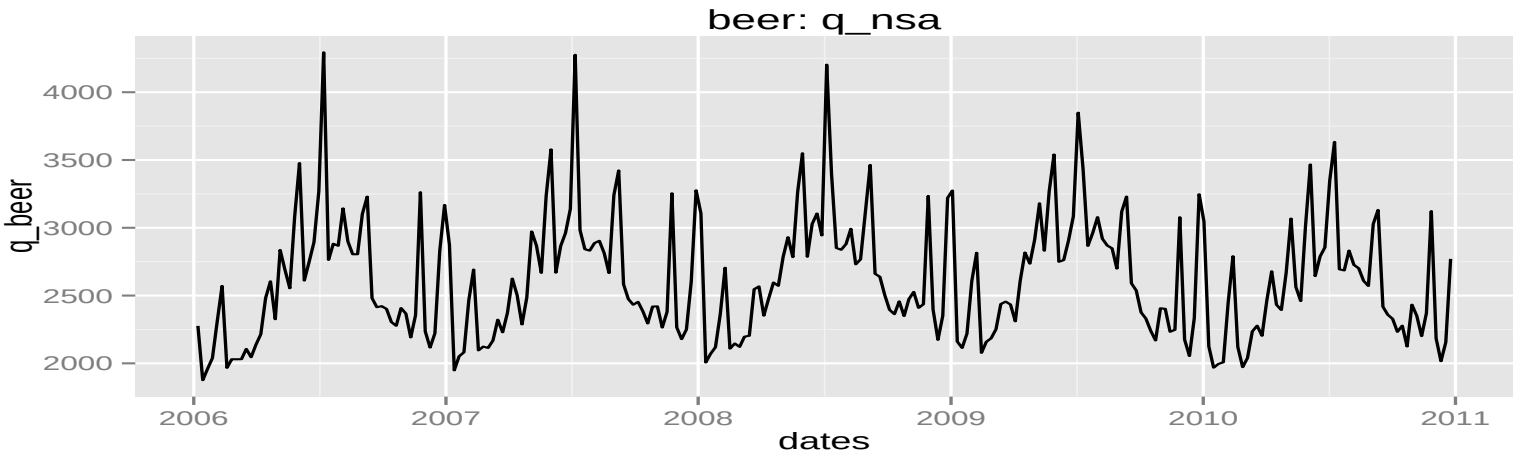
- I have importance sampling distribution for P, Q, V.
- But P, Q and V are not independent. How do I sample them jointly?
- some thoughts to select columns:
  1. Vertically stack P, Q, V into a $3T \times N$ matrix A:
  2. Select N_p columns from P, N_q from Q, N_v from V. Combine columns.
- What are the properties?

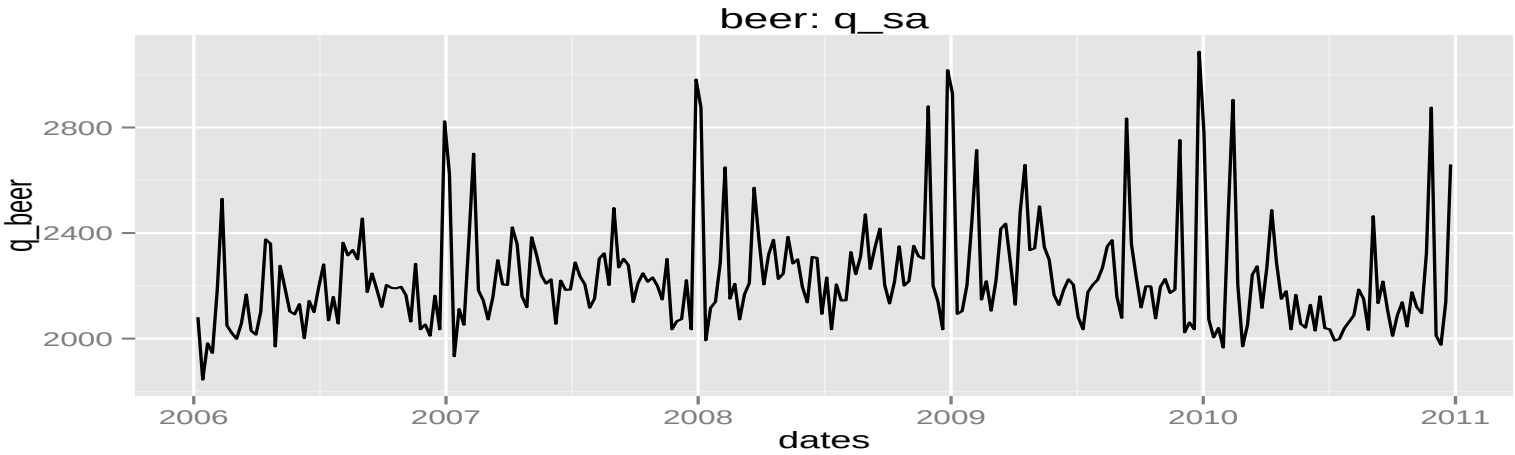
Issue 3: accommodate prior information

- Numerical algorithms approximate A
- we may want stores from all regions in the country, with a mix of big and small firms.
- Numerical sampling could favor big states (New York, California) and big stores (Walmart).
- How to do CSSP subject to certain constraints on the distribution, (e.g consistent with spatial and size distribution of stores.)

Issue 4: data irregularities

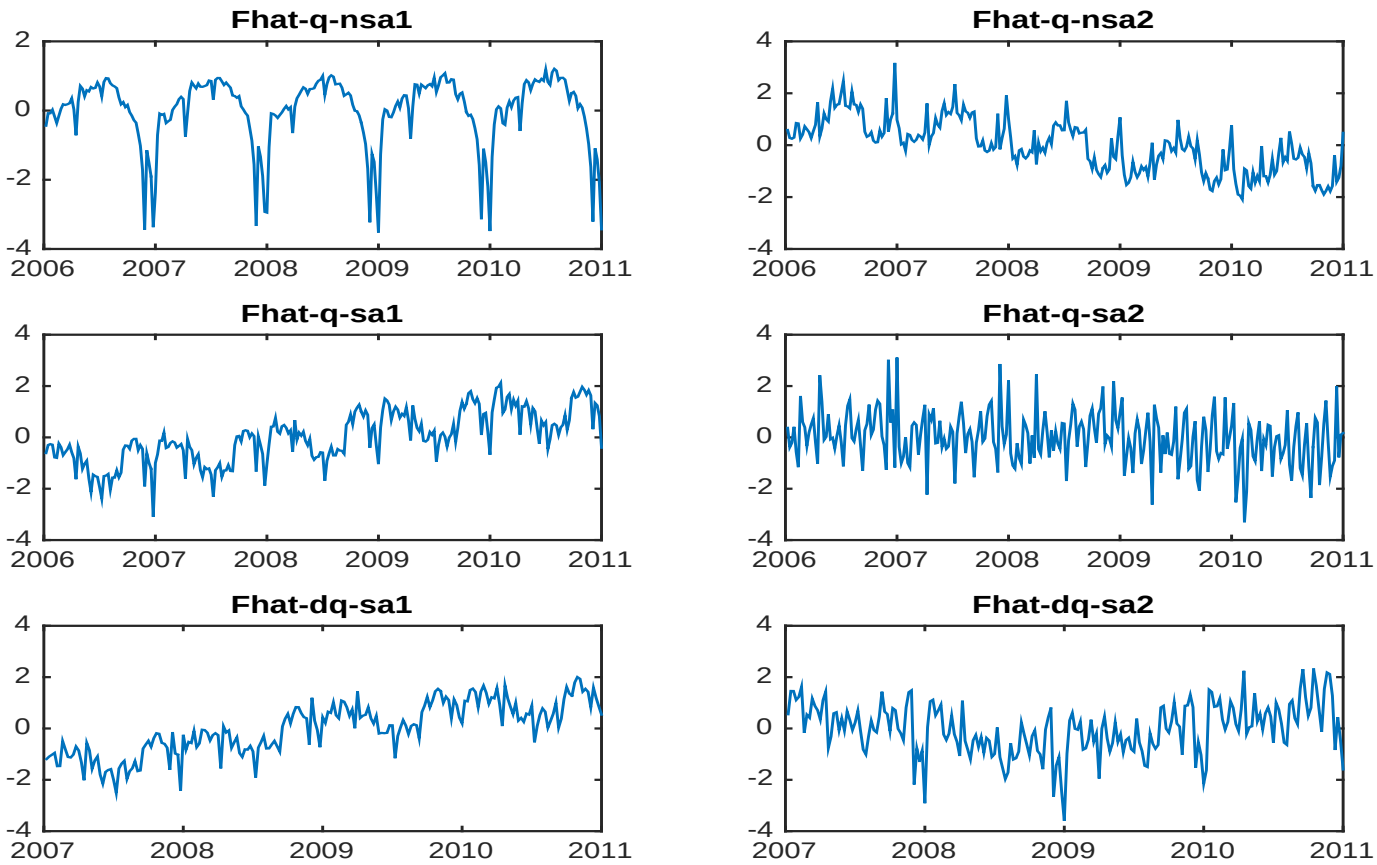
- Official Data: clean, integrity checked.
- Internet data: noisy, revised.
- Scanner data: weekly data, seasonally unadjusted,
  - Holiday effects: beer in summer, wine at christmas.
  - Removing seasonal effects not a trivial problem even for 1 series. We have billions of series. Recall:
    $y_{ti} = \underbrace{T_{ti}}_{\text{trend}} + \underbrace{C_{ti}}_{\text{cycle}} + \underbrace{S_{ti}}_{\text{seasonal}} + \underbrace{I_{ti}}_{\text{irregular}}$.
    Cannot analyze the cycle without taking care of the seasonal and irregular variations.





- Extract PCA by first removing S_{it} from y_{it}.
- Failed! PCA still have strong seasonal variations.
- Idea: Treat seasonal variations as low rank variations. Let RPCA remove them directly.

- weekly data, aggregated by upc, location. A: 260×103.



- Top panel: PCA raw data.
- Second panel: PCA of $y_{it} - S_{it} = F_t$.
- Third panel: Take $F_t - F_{t-52}$.

- Can we use RPCA? is coherence condition satisfied?

$$r = O\left(\frac{\log \min(T, N)}{(\log \max(T, N))^2}\right).$$

- $T = 60, N = 1071$: $r = 1$.
- $T = 260, N = 105$: $r \approx 3$.
- More than 3 seasonal components alone!

Concluding Remarks

- Problem: data highly heterogeneous (location, store size).
- The heterogeneity that makes the data interesting also make them hard to work with.
- Need to train algorithmw to extract low rank components from less structured data.
- A not so small challenge: new concepts and terminology! Map algorithmic results to statistical concepts.
- Interdisciplinary work. Welcome collaboration.