Exploring New Data Sources with New Methods

Serena Ng

June 2018

Columbia University and NBER

Collaborators: Rishab Guha (Harvard) and Evan Munro (Stanford)

Conventional Macroeconomic Data

- Intentionally collected:
  - Clean and checked
  - Compact and regular.
- Seasonally adjusted: X13, SEATS/TRAMO (parametric).
- Mostly metrical (not qualitative) variables.

New Sources

- Not always intentionally generated. By-product of economic and social activities.
- Unconventional characteristics
  - 3V: Volume, Variety, Velocity.
  - Can be non-metrical and not even numerical.
  - Time series analysis not appropriate for survey and text data.
  - Sample size: lots of series, but short span.
  - Large sample theory?
- Available as is: pre-processing is responsibility of user.

This Talk

- We have been examining old data using new methods.
- Let's explore new data! Not surprisingly, new challenges.
- Share my experience with scanner data and survey responses.
- Use economic theory, econometrics, machine learning tools.

The Nielsen Scanner Data, 2006-2014

- The good:
  - N_p=1000+ products from N_g=100+ product groups,
  - $T = 469$ weeks, 2006:1:07 -2014:12:29. Financial Crisis
  - actual sales, not estimates from surveys.
  - observe transactions price, not price indices.
  - weekly (not monthly/quarterly), major MSA (not 4 regions).
- The not so good for macro analysis:
  - groceries, mass merchandise products: few durable goods.
  - $N_{\text{years}} = 9$, short span.
  - multi-dimension heterogeneity.

What level of Aggregation?

- Ng (2017): work at (store,upc) level.
  - (p, q, sales) in balanced panels, 1000+ products
- This paper:
  - sales product groups (g), counties (c) in states (s)
  - Each (g,c) data matrix is $T = 469$ by $N_g = 108$.
  - s: NY, CA, TX, FL.
- Goal: Find useful macroeconomic information from data.
  - Need a way to summarize broad-based information.
  - Need a way to separate seasonal from cyclical variations.

Demand System: Theory

- The rank of a demand system is the maximum number of functions F spanned by (p, income).
- rank 1: share$= F_1(\text{income})\Lambda_1'$
- rank 2: share $= F(p, \text{income})\Lambda'$.
  - e.g. Linear expenditure system, translog
  - PIGLOG class, AIDS (Deaton-Muellbauer).
  - a big literature pre-BLP.

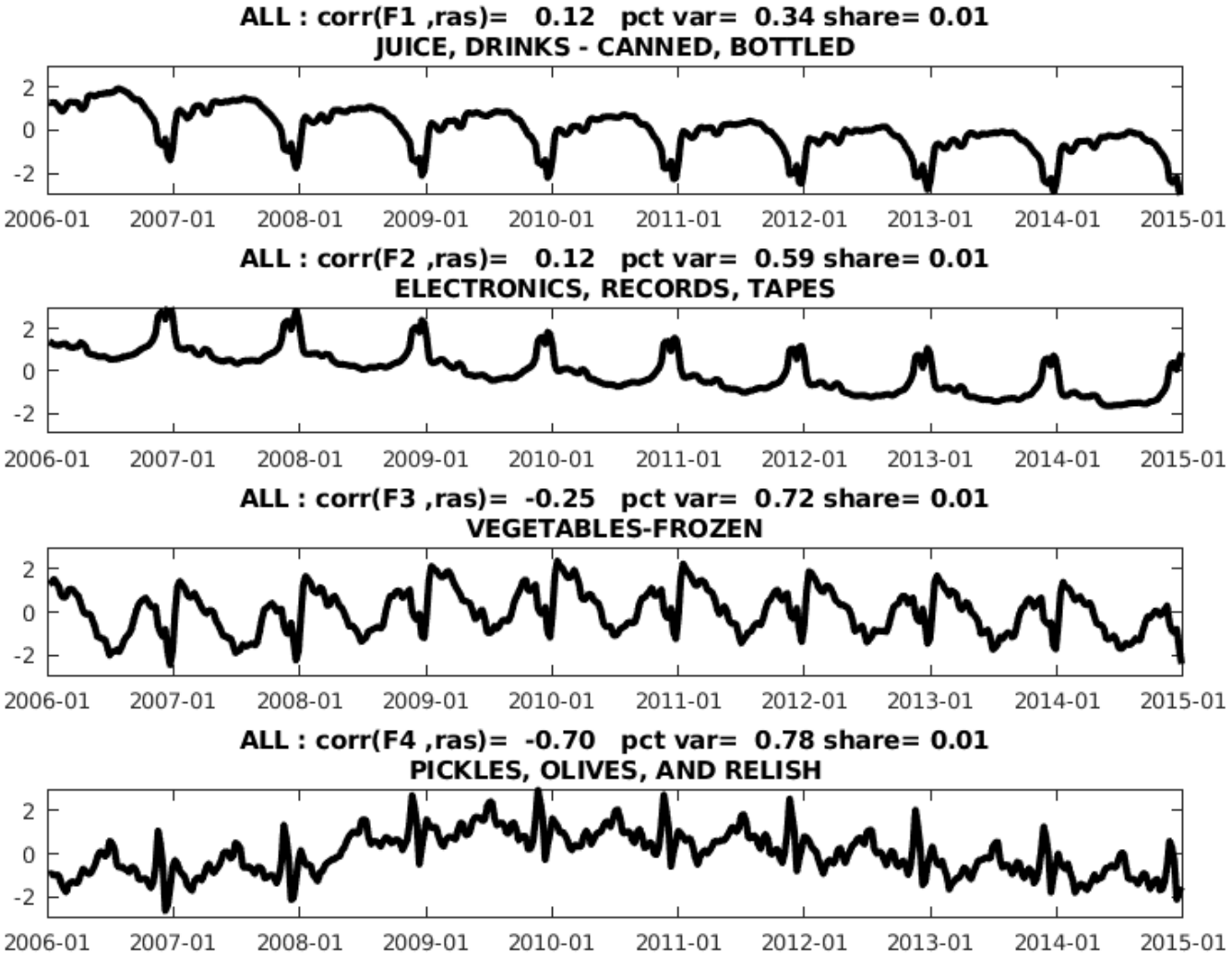
Demand System: Estimation

$$\text{share} = p^*(\Psi)\Lambda_1' + \text{income}\ \Lambda_2' + e.$$

- Classical estimation: small number of product groups, N_g.
  - cross-section analysis: many households, one or few years.
  - time series analysis: many years, average consumer.
  - use economic assumptions to construct p^*. Linear model.
  - estimate of rank: between 2 and 4.

- Nielsen: each c in s, data matrix is 469×108.
  - can consistently estimate F and Λ from sales data
    - ie. without using price/income data
  - non-parametric in economic and econometric sense.
  - can also consistently estimate the rank of demand system, r.
  - In Nielsen data, rank > 5. Why?

Factors Estimated from Data NSA: All



Strong seasonality!

Seasonality

- 3 challenges
  - i spending is concentrated in the last 6 weeks of year.
    - in demand system, income=pq and q is seasonal.
    - entry-exit is seasonal: more goods introduced in Q4.
  - ii weekly data: not exactly periodic (Gregorian calendar).
  - iii short span: $T = 469$, but $N_{\text{years}} = 9$.
- Need automated and scalable seasonal adjustment method.
- We treat seasonal adjustment as a prediction problem.
- Two-step panel procedure. Remove seasonality one year at a time, rather than one series at a time.

Seasonal Adjustment as a Prediction Problem

	county	prod. group	week	year	state
	c	g	t	$\tau = yr(t)$	s
value	330	108	469	9	4

Step 0: remove size effect by within year demeaning.

For each product group g and county c, standardize by year:

$$
\begin{aligned}
y_{gct} &\equiv \frac{\log(\text{SALES}_{gct}) - \mu_{gc\tau}}{\sigma_{pc\tau}} \\
&= \underbrace{d_{gct}}_{\substack{\text{specific, seasonal,} \\ \text{excatly periodic}}} + \underbrace{q_{gct}}_{\substack{\text{common, seasonal,} \\ \text{not exactly periodic}}} + \underbrace{u_{gct}}_{\text{cyclical}} .
\end{aligned}
$$

Step 1: Specific seasonality: $d_{gct} = \alpha^0_{gc} + \text{Fourier}(t; \beta_{gc})$

$$y_{gct} - \widehat{d}_{gct} = \widehat{q_{gct} + u_{gct}}.$$

- Deterministic and smooth, Cleveland et al (2007, 2014).
- One time series regression for each (g, c).

Step 2: Common seasonality: estimate q_{gct} from $\widehat{q_{gct} + u_{gct}}$

- pool information across heterogeneous counties.
- Train algorithms to do prediction.

Step 3: Rescaling:

$$\begin{aligned} y_{gct} &= \alpha_{g0} + \alpha_{g1} \cdot \widehat{d}_{gct} + \alpha_{p2} \cdot \widehat{q}_{gct} + u_{gct} \\ x_{gct} &= \widehat{\text{log sales}}^{sa}_{gct} \equiv \widehat{u}_{gct} \cdot \sigma_{g\tau} + \mu_{g\tau} \end{aligned}$$

Step 2 in More Detail $y_{tct} = d_{gct} + q_{gct} + u_{gct}$.

Key observation: q_{gct} is common and predictable.

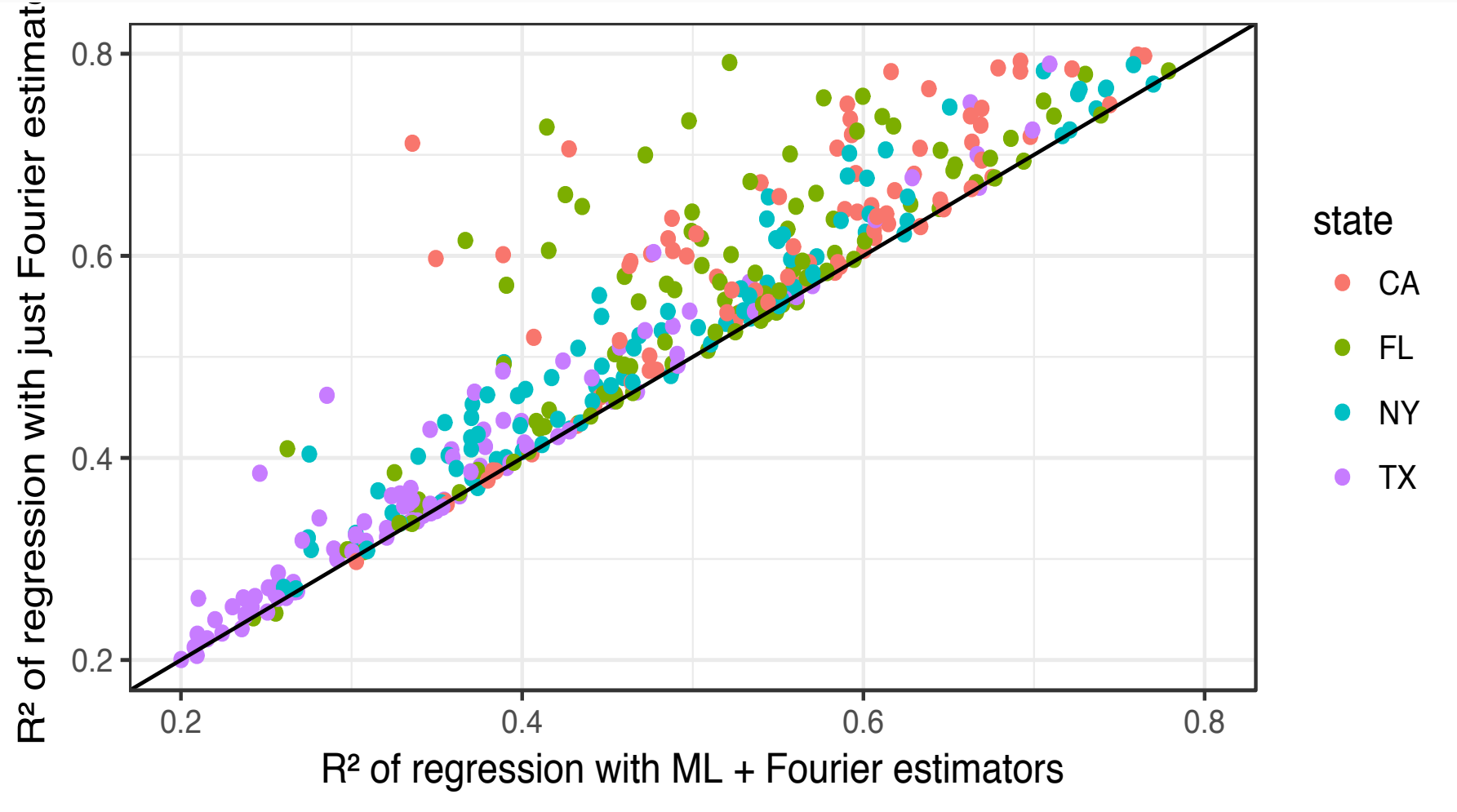
$$\widehat{q_{gct} + u_{gct}} = \psi_{gc} Z_{gct} + \text{err}_{gct}.$$

- pool information across counties and over years.
- Z_{gct}: predictors
  - i county specific: social-economic, weather and location.
  - ii day-specific: holidays, sports events, back to school.
  - iii interaction of (i) and (ii).
- $\approx$ 400 dummy predictors problem for each (g, τ) panel.

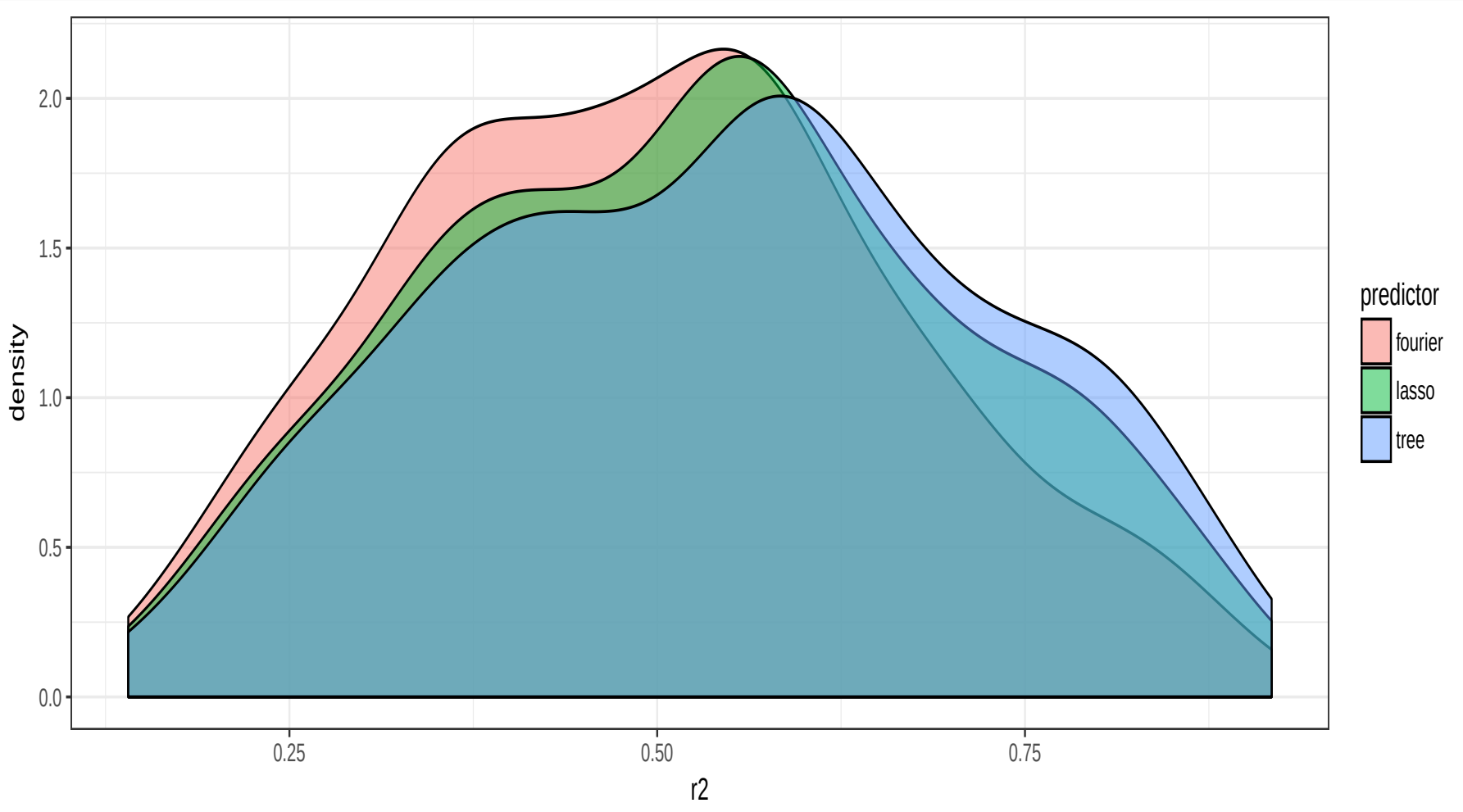
Model Choices

- Many predictors: regularization (lasso, boosting).
- Unknown function form, possibly non- smooth (spikes).
- Regression tree: non-parametric, but high variance.
- Random forest: flexible, non-smooth, ensemble averaging.
- Data $\mathcal{D}^s_{g\tau} = (\mathcal{D}^s_{1,g\tau}, \mathcal{D}^s_{2,g\tau})$=(training, prediction)
  - ncol$(\mathcal{D}^s_{g\tau}) =$ # of predictors+1.
  - Given (τ, s), rows of $\mathcal{D}^s_{1,g\tau} = (469 -$ weeks in $\tau) \times N^s_c$.

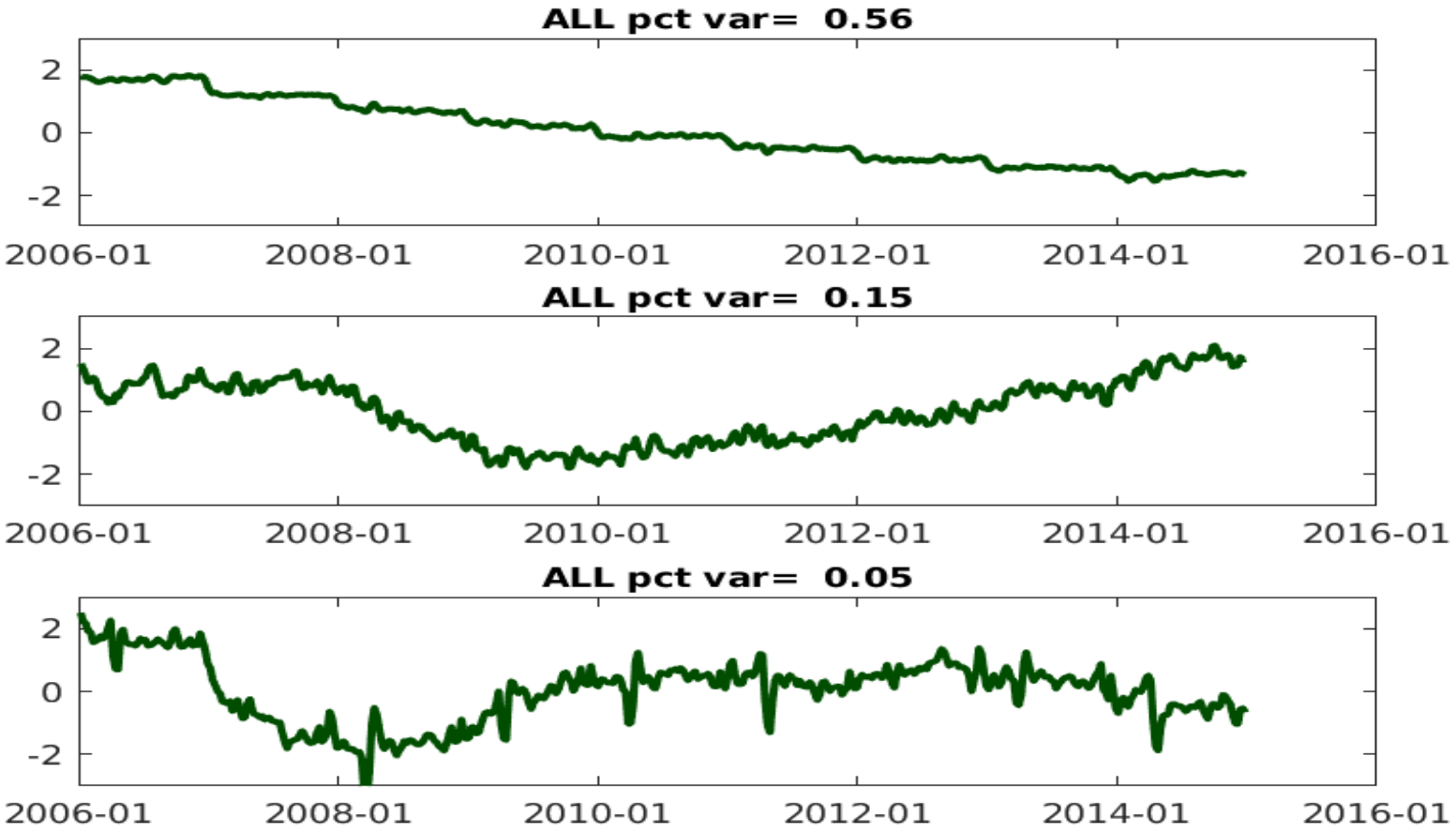
Improvement over Fourier regression



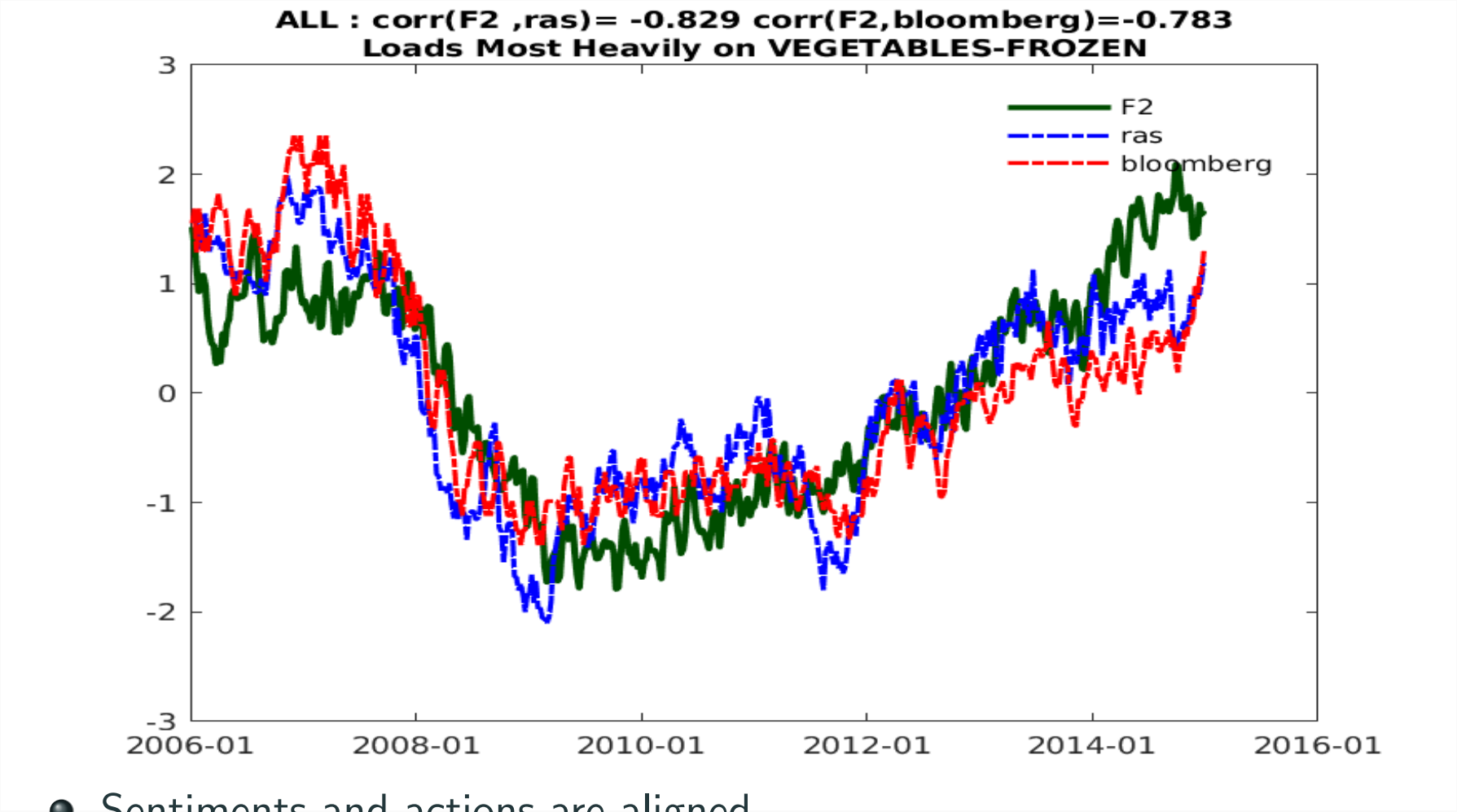
Comparison with Fourier and Lasso



Factors Estimated from Data SA: ALL



$\widehat{F}_2$, Rasmussen, Bloomberg



- Sentiments and actions are aligned..
- Cyclical factor loads heavily on pasta and frozen vegetables.
- Accounts for about 15% of seasonally adjusted data.

1. Nielsen Scanner Data

Nielsen Scanner Data

Budget Shares and Demand System

Seasonal Adjustment as a Prediction Problem

2. Non-metrical Data

From Modeling Topics to Modeling Surveys

Data Types

- Continuous data, e.g stocks and flows.
- Discrete data, e.g. socioeconomic status indicators.
  - non-metrical
    - binary (yes/no)
    - nominal (unordered data such as zip code)
    - ordinal (approval ratings, consumer sentiments)
  - metrical: count (age, income class)

Survey Responses

- Michigan Survey of Consumer Sentiment (monthly)
- Bloomberg Survey of Consumer Comfort (weekly)
- Conference Board's Consumer Confidence Index (monthly)
- Rasmussen Survey of investors/non-investors (daily)
- Gallop survey (weekly).

Example: Michigan Consumer Sentiment Survey

1. Would you say that you are better off or worse off financially than you were a year ago?
2. Now looking ahead–do you think that a year from now you will be better off financially, or worse off?
3. Now turning to business conditions in the country as a whole–do you think that during the next 12 months we'll have good times financially?
4. Looking ahead, which would you say is more likely –that in the country as a whole we'll have continuous good times during the next five years or so, or that we will have period of widespread unemployment or depression, or what?
5. Generally speaking, do you think now is a good or bad time for people to buy major household items?

Michigan data

(1)	(2)	(3)	(4)	(5)
PAGO	PEXP	BUS12	BUS5	DUR
1 better now	1 will be better	1 good times	1 good times	1 good
3 same	3 same	2 tood w. qual	2 good w qual	3 pro-con
5 worse	5 will be worse	3 pro-con	3 pro-con	5 Bad
8 DK	8 DK	4 bad w. qual	4 bad w qual	8 DK
9 NA	9 NA	5 bad times	5 bad times	9 NA
		8 DK	8 DK	
		9 NA	9 NA	

UK Inflation Attitudes

1. How have prices have changed over the last twelve months?
   - *1 (down+) - 8 (up+), 9 (dk)*
2. How much will prices change over the next twelve months?
   - *1 (down+) - 8 (up+), 9 (dk)*
3. If prices started to rise faster, you think Britain's economy would be?
   - *1 (stronger) - 3: weaker, 4: (dk)*
4. The government has set an inflation target of 2%...Is
   - *1(high), 2(low), 3(about right), 4 (dk)*
5. Overall, how satisfied or dissatisfied are you with the way the Bank of England is doing its job to set interest rates in order to control inflation?
   - *1 (satisfied) - 5 (dissatisfied), 6 (dk)*

Small Literature

How to find patters in survey data?

- ordinal data: standardize, do PCA.
- polychloric correlations. parametric
- Filmer-Prichett method: treat each option in each question as a binary variable. Do PCA.
- Impute continuous data from discrete data (psychology).

Topic Models

View a document is a mixture of K topics.

- D documents, vocabulary of size V.
- Words in document d :$w_d = \{w_{d,1}, \ldots, w_{d,V}\}$.
- Observe $X \in \mathbb{R}^{D \times V}$: matrix of frequency of word v in doc d.
    - eg. economic topic has words 'inflation','unemployment'.

Goal: uncover topics in collection of documents.

X: a $D \times V$ word-document co-occurence matrix

1. Matrix factorization $X \approx \mathbb{U}_K \mathbb{D}_K \mathbb{V}_K^T = \theta\beta'$.
   - β and θ have rank $K < \min(D, V)$.

2 Statistical model: p(word)=p(word|topics)×p(topics|doc).
   - Topics are generated before words and doc.
   - Specify multinomial distribution for topics and word.
   - Dirichlet distribution is conjugate prior for multinomial.

Likelihood: $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$

$$\prod_{k=1}^{K} \underbrace{p(\beta_k)}_{\text{Dirichlet};\eta} \prod_{d=1}^{D} \underbrace{p(\theta_d)}_{\text{Dirichlet};\alpha} \times \left(\prod_{v=1}^{V} \underbrace{p(z_{d,v}|\theta_d)}_{\substack{\text{Topic assignment} \\ \text{depends on} \\ \text{topic proportions}}} \underbrace{p(w_{d,v}|\beta_{1:K}, z_{d,v})}_{\substack{\text{Word depends on} \\ \text{topic assignment} \\ \text{and all topics.}}} \right).$$

Survey Data

Assume: responses reflect time-varying mixture of K sentiments.

- Survey question j has L_j distinct categorical answers.
- Total permutation of responses: $V = \prod_{j=1}^{J} L_j$.
- N respondents grouped into T groups of N_t.
- X_{tv} is $\#$ responses to permutation v from survey at time t.

Goal: recover the K sentiments from data X.

Mapping Topic Modeling Into Survey Modeling

Topic Modeling		Survey Modeling	
$\theta^{D\times K}$	doc-topic mixture	$G^{T\times K}$	time-factor mixture
$\beta^{K\times V}$	word-topic prob.	$\Lambda^{K\times V}$	factor-response prob.
D	documents	T	time periods
V	# words	V	# permutations of responses
v	index vocabulary	v	index response permutation
N_d	word counts in doc d	N_t	# response at t
$X_{d,v}$	word count	$X_{t,v}$	permutation counts
$w_{d,n}$	word n in doc d	$s_{t,i}$	respondent i at time t

Generative Model For Survey Modeling

1. Draw $V \times 1$ vector $\Lambda_{k,:}$ from $\mathcal{D}(\eta)$.
2. Draw $K \times 1$ vector $G_{t,:}$ from $\mathcal{D}(\alpha)$
3. for $i \in N_t$, $t = 1, \ldots, T$:
   a. draw sentiment assignment $z_{t,i} \in [1, K]$ from G_t
   b. draw response $S_{t,i}$ from $\Lambda_{z_{t,i}}$.

   Aggregation: $X_{t,v} = \sum_{i=1}^{N_t} 1(S_{t,i} = v)$.
4. Likelihood: $p(\Lambda_{1:K,:}, G_{1:T,:}, Z_{1:T,1:N}, S_{1:T,1:N}) =$

$$\prod_{k=1}^{K} \underbrace{p(\Lambda_{k,:})}_{\text{Dirichlet};\eta} \prod_{t=1}^{T} \underbrace{p(G_{t,:})}_{\text{Dirichlet};\alpha} \times \left(\prod_{t=1}^{T} \prod_{i=1}^{N_t} \underbrace{p(Z_{t,i}|G_{t,:})}_{\substack{\text{sentiment} \\ \text{assignment} \\ \text{depends on} \\ \text{response} \\ \text{proportions}}} \quad \underbrace{p(S_{t,i}|\Lambda_{1:K,:}, Z_{t,i})}_{\substack{\text{response depends on} \\ \text{sentiment assignment} \\ \text{and all responses.}}} \right).$$

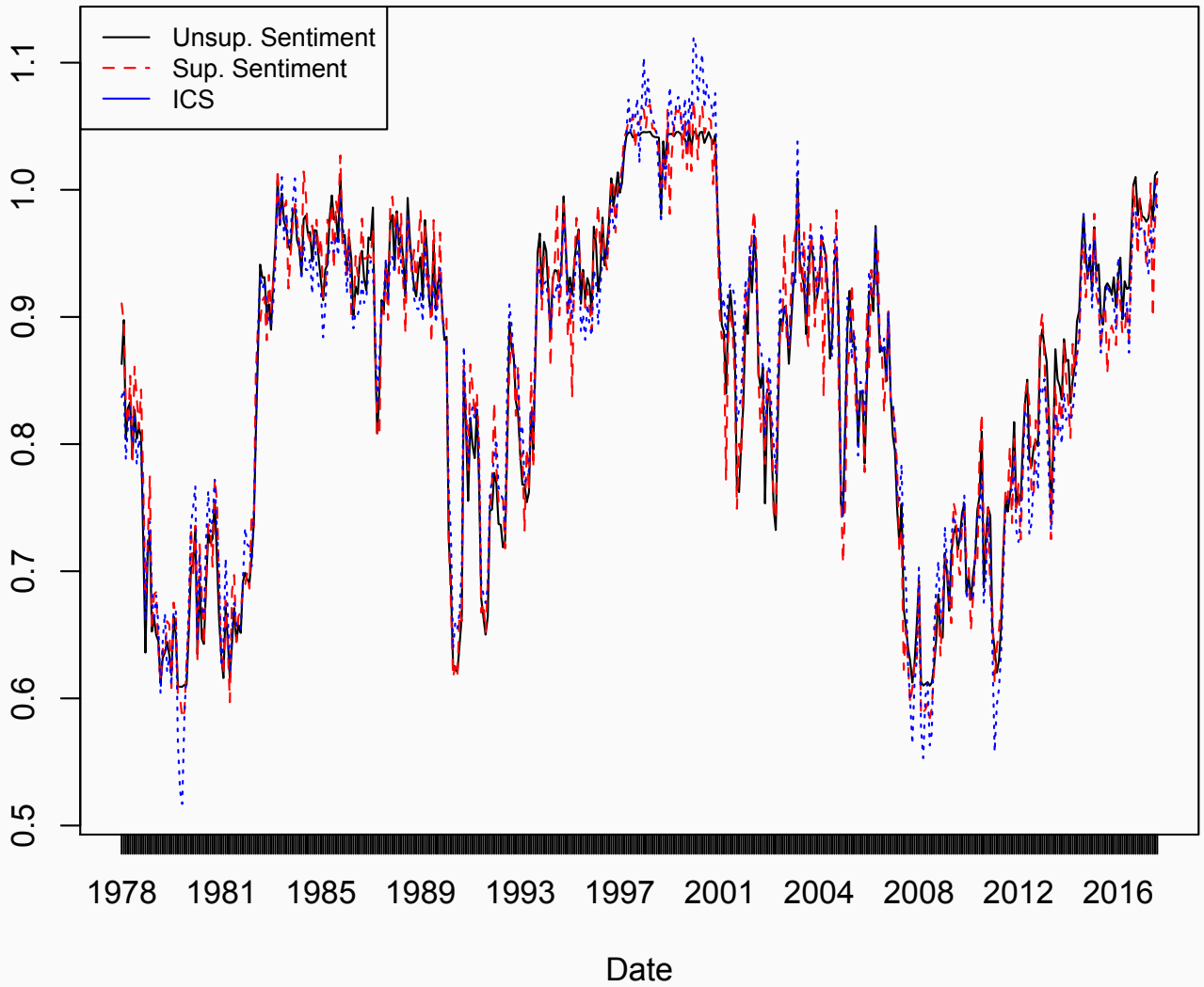
Estimated Sentiments: Michigan Data

- 1: Better/Good, 3: Same/Even, 5: Worse/ Bad

Responses with the highest prob. for the two sentiments

Positive Sentiment	Negative Sentiment
11111	53551
13111	53555
31111	55555
51111	55551
13151	13551

Estimated Positive Sentiment Index vs. Michigan



Inflation Forecasting

Michigan ICS:

$$\text{infl}_{t+1} = \beta_0 + \text{infl}_t + \text{michigan}_t\beta_1 + \epsilon_{t+1}$$

Unsupervised LDA:

$$\text{infl}_{t+1} = \text{infl}_t + g_t\beta + \epsilon_{t+1}$$

Supervised LDA (include forecasted variable in likelihood):

$$\text{infl}_{t+1} = \text{infl}_t + \overline{z}_t\beta + \epsilon_{t+1}.$$

Out of Sample MSE

AR(1)	michigan	unsupervised	supervised
0.0505	0.0430	0.0430	0.0427

- Approach allows
  - better understand how each sentiment affects reported index.
  - predictive distribution, uesful for forecasting.
  - treatment of missing values and non-response.
- Work to do
  - introduce dynamics (Markov switching)
  - identification: LT/diagonal matrix, anchor words.
  - how to determine K?
  - alternatives to MCMC.

Conclusion

- Conventional macro modeling: national accounts data.
- Potential gains from exploring new data, but challenges.
- Integrate modern methods with economic/ecnometric theory.
- Data Issues and Replicability