A MACHINE LEARNING ANALYSIS OF SEASONAL AND CYCLICAL SALES IN WEEKLY SCANNER DATA

July 13, 2018

Rishab Guha Harvard University
Serena Ng Columbia University and NBER

Macroeconomic Data

- Conventional data
  - Intentionally collected
    - Cleaned, adjusted, and checked
    - Compact and regular.
    - Mostly metrical (not qualitative) variables.
- New Data
  - Often by-product of economic and social activities.
  - Unconventional characteristics:
    - 3V: Volume, Variety, Velocity.
    - Can be non-metrical and not even numerical (eg. text).
    - Available as is: pre-processing is responsibility of user.

The Nielsen Scanner Data

The good:

- Volume (5TB):
- over 1000 products from N_g ≈110 product groups,
- Actual sales, not estimates from surveys.
- 2006:1:07 -2014:12:29. Financial Crisis
- Weekly frequency: (not monthly/quarterly)
- Local level economic data: major MSA (not 4 regions).

The not so good especially for macro analysis:

- Groceries, mass merchandise products: few durable goods.
- $T = 469$ weeks, $N_{\text{years}} = 9$, short span.
- Volume: 5TB, memory constraint.
- Variety (multilevel heterogeneity): product, spatial, time:
- Velocity: weekly, quasi-periodic.
  - Few studies take advantage of weekly frequency.
  - A Challenge: Strong Seasonality.

Outline

1. Motivation
2. Demand Analysis Pre-adjustment
3. Seasonal Adjustment
4. Demand Analysis Post-adjustment
5. Cyclical Sensitivity

Work in Progress: Comments Welcome

Disclaimer

- Calculated (or Derived) based on data from The Nielsen Company (US), LLC and marketing databases provided by the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business.
- The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

Which Variables?

- Ng (2017): price and quantity at (store,upc) level.
    - big data= big time cleaning data.
- Guha and Ng (2018): sales.
    - What data say about small purchases especially around 2008?
    - Is there value in monitoring this weekly data?
    - extract information at county and group levels.
    - weekly frequency.

	county	**group**	**week**	**year**
index	$c(s)$	g	t	τ
total	N_c	$N_g = 108$	$T = 469$	$N_{\text{yr}} = 9$

Budget Shares: Most Purchased Categories

CA: $N_c = 53$		FL: $N_c = 58$		NY: $N_c = 58$		TX: $N_c = 161$	
3.4	bread	4.4	medications	4.1	medications	3.7	carbon. bev
3.3	beer	4.3	tobacco	3.2	fresh prod.	3.7	medications
3.3	juice	3.1	carbon. bev.	3.1	bread	3.4	snacks
3.2	wine	2.9	liquor	3.0	candy	2.9	bread
3.0	fresh prod.	2.8	beer	2.8	snacks	2.8	tobacco
3.0	carbon. bev	2.6	juice	2.8	juice	2.6	pkgd meat
3.0	snacks	2.6	candy	2.7	tobacco	2.6	candy
2.7	pkgd meat	2.4	snacks	2.5	beer	2.5	fresh prod.
2.7	salad dress.	2.3	milk	2.4	carbon. bev	2.5	juice
2.6	medication	2.6	bread	2.3	milk	2.5	beer

Guiding Framework

- A demand system expresses expenditures or shares as functions of prices $p = (p_1, \ldots, p_{N_g})$ and income.

  $$\text{share}_g = \sum_{i=1}^{r} \lambda_{ig}(\log \mathrm{p}) F_i(\log \mathrm{p}, \log(\text{income}/P)).$$

  P is a theory based price index, F_1 constant for adding up.

- The rank of a demand system r is the dimension of F.
  - rank 1: homothetic demand independent of income.
  - rank 2: demands are generalized linear (Muellbauer 1980).

- Demand theory forms basis of consumer price indexes.

Empirical Demand Systems

- Approximate spending by flexible functions. Under restrictions of consumer theory. eg. LES, translog.

$$\text{share}_g = \lambda_{0g} + \sum \lambda_{jg} \log p_g + \beta_g \log(\text{income}/P) \quad \text{(AIDS)}$$

  - p and income/P are common across groups g.
  - P imposes cross-equation restrictions. Use proxy simplifies.

    $$\text{share}_g = \lambda_{0g} + \sum \lambda_{jg} \log p_g + \beta_g \log(\text{income}/P^*) + error_g.$$

    In AIDS, $\log P^* = \sum_j w_j \log p_j$ is Stone's price index.
  - $error_g$: omitted time varying preferences or measurement error (less likely in scanner data).

Empirical Estimate of r

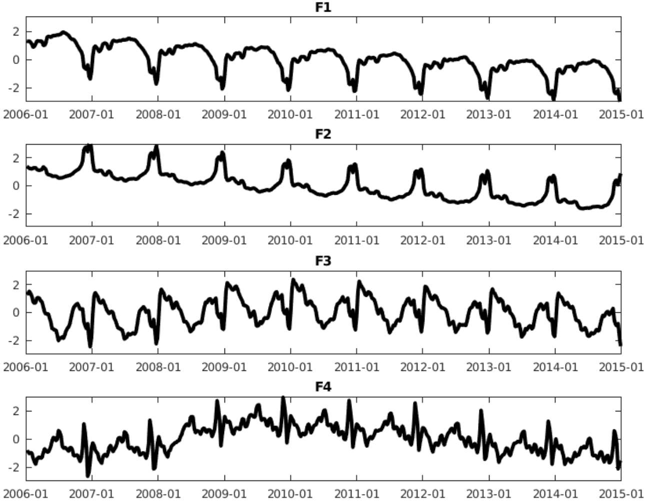
- Classical estimation: T large, N_g small.
  - cross-section analysis: T households, one or few years.
  - time series analysis: T years, average consumer.
  - typical estimate using seasonally adjusted data: $2 \leq r \leq 4$

- Nielsen data: $(T = 469) \times (N_g = 108)$ for each s.
  - budget shares have a factor structure.

$$\text{share}_{gt} = F_t' \Lambda_g + e_{gt}.$$

  - estimate latent F and Λ without using price/income data
  - non-parametric in economic and econometric sense.
  - rank of demand system in Nielsen data > 5. Why?

Factors Estimated from Raw Data: All States



Strong and heterogeneous seasonal effects! Where is the cycle?

Dealing with Seasonality

- Spending is concentrated in the last 6 weeks of year.
    - 2 stage budgeting: income$=\sum_g p_g q_g$ is seasonal.
    - entry-exit is seasonal: more goods introduced in Q4.
- 3 challenges in deseasonalization
    - i weekly data: not exactly periodic, (Gregorian calendar).
        - earliest Easter: March 23, 2008, latest easter, April 24, 2011.
        - beer sales depends on whether July 4th is thursday or sunday.
    - ii volume and heterogeneity.
        - effects of holiday events differ by regions and products.
    - iii short span: $T = 469$, but $N_{\text{years}} = 9$.
- 52 week differencing does not work.

Bottom-Up using Time Series Methods?

- Univariate (parametric) procedures: X13, SEATS/TRAMO: developed for exactly periodic data, large sample theory.
- Let $y_{gt} = \log(\text{sales}^{s}_{gt})$ on group g in some given s.
- Let $\widehat{seas}_{gt}$ be some univariate estimate:

$$\begin{aligned}\widehat{\text{cycle}}_{gt} &= y_{gt} - \widehat{\text{seas}}_{gt} = \text{cycle}_{gt} + \text{seas}_{gt} - \widehat{\text{seas}}_{gt} \\ &= \text{cycle}c_{gt} + (1-\theta_g)\text{seas}_{gt}.\end{aligned}$$

Suppose for each g, $\theta_g = \text{corr}(\text{seas}_{gt}, \widehat{\text{seas}}_{gt}) \approx 1$.

Common Seasonality ⇒ Common Adjustment Error

$$\begin{aligned}
y_{gt} &= \text{cycle}_{gt} + \text{seas}_{gt} \\
\text{seas}_{gt} &= \lambda_g^{seas\prime} F_t^{seas} + e_{gt}^{seas} \\
\text{cycle}_{gt} &= \lambda_g^{cycle\prime} F_t^{cycle} + e_{gt}^{cycle}.
\end{aligned}$$

- Target variable: cycle_t

$$\begin{aligned}
\widehat{\text{cycle}_t} &= \frac{1}{N_g}\sum_{g=1}^{N_g} \text{cycle}_{ct} + \frac{1}{N_g}\sum_{g=1}^{N_g}(1-\theta_g)\text{seas}_{ct} \\
&= \overline{\text{cycle}}_t + \frac{1}{N_g}\sum_{g=1}^{N_g}(1-\theta_g)\left(\lambda_g^{seas\prime} F_t^{seas} + e_{gt}^{seas}\right) \\
&= \overline{\text{cycle}}_t + \overline{(1-\theta_g)\lambda_g^{seas\prime}} F_t^{seas} + o_p(1).
\end{aligned}$$

- unless filter is ideal or there is no common seasonality, (ie. $\theta_g = 1$ or $\lambda_g^{seas} = 0$) $\forall c$, bottom up leaves F_t^{seas} in estimate of cycle.

- Ng (2017)

Perfect Univariate Filtering Unlikely

- Expect common seasonality in estimate of common cycle extracted from data adjusted on a series by series basis.

  i Spikes from holiday sales not fixed across years.

  ii Smooth functions are not good at picking up spikes.

  iii Span of data is short. Finite sample bias.

- Need some way to remove the common seasonal variations that univariate methods fail to capture.

A New Approach

- Given goal of understanding cyclical patterns, we remove common seasonality year-by-year using a panel approach. This complements the univariate adjustments.

- Since weekly seasonal effects are quasi-periodic, we use a periodicity-free definition of seasonality.

- Motivation: sales of group g in neighboring counties have similar seasonal patterns regardless of county size.

A Prediction-Based Approach to Seasonality

- For each (c, g), standardize year by year, remove size effect:

$$y_{gct} \equiv \frac{\log(\textsc{sales}_{gct}) - \mu_{gc\tau}}{\sigma_{gc\tau}} \quad t \in \tau.$$

$$\begin{aligned} y_{gct} &= d_{gct} + q_{gct} + u_{gct} \\ &= \underbrace{\text{specific seasonal} + \text{correlated seasonal}}_{\text{seasonal}} + \text{cyclical}. \end{aligned}$$

- Key: seasonal components d_{gct} and q_{gct} are predictable.

- Treat seasonal adjustment as a prediction problem.

Overview: $y_{gct} = d_{gct} + q_{gct} + u_{gct}$

- Step 1: Fourier regression: y_{gct} on $d_{gct} \Rightarrow$ resids $\widehat{q_{gct} + u}_{gct}$

$$d_{gct} = \sum_{j=1}^{p_d} \beta^s_{k,2j-1} \mathrm{sin}(\delta_{tj}) + \beta^s_{k,2j} \mathrm{cos}(\delta_{tj}) + \sum_{j=1}^{p_m} \psi^s_{r,2j-1} \mathrm{sin}(m_{tj}) + \psi^s_{k,2j} \mathrm{cos}(m_{tj}$$

where $\delta_{tj} = 2\pi j \frac{\text{day of year}_t}{\text{days in year}}$ and $m_{tj} = 2\pi j \frac{\text{day of month}_t}{\text{days in month}}$. Cleveland (1984).

- Step 2: pool information across counties and years:
    - predict q_{gct} using machine learning algorithms.

- Step 3: Reweight and rescale:

$$\begin{aligned} y_{gct} &= \alpha_{p0} + \alpha_{g1} \cdot \widehat{d}_{gct} + \alpha_{g2} \cdot \widehat{q}_{gct} + u_{gct} \\ x_{gct} &= \widehat{\text{log adjusted sales}}_{gct} \equiv \widehat{u}_{gct} \cdot \sigma_{g\tau} + \mu_{g\tau} \end{aligned}$$

Step 2: Intuition

- For each (g, s, τ): predict q_{gct} based on $\widehat{q_{gct} + u_{gct}}$ $\forall c \in s$.
- Control for multidimensional sesaonal heterogeneity using lots of dummy predictors using a flexible seasonality function.
    - Intuition from LSDV: incidental parameter if T is short.
    - Fok, Franses, Paap (2007): hierarchical structure, Bayesian.
    - We let algorithms choose predictors and functional form.

Step 2: Predictors $\mathbb{Z}_{gct}$

- Many (391) predictors

  i (base set) date-specific dummies: holidays, sports events.

  ii social-economic indicators at county level.

  iii weather and location from NOAA.

  iv interaction of (i) and (ii).

- Data $\mathcal{D}_{g\tau} = (Y_{g\tau}, Z_{g\tau})$=(response,predictors)
  $(\mathcal{D}_{1,g\tau}, \mathcal{D}_{2,g\tau})$=(training, validation)

- ncol($\mathcal{D}_{g\tau}$) $=$ # of predictors$+1$.

- nrow $(\mathcal{D}_{1,g\tau}) = (469 -$ weeks in year $\tau) \times N_c^s$.

step 2: Machine Learning

a. Pooled OLS, non-regularized, no averaging.

b. LARS type algorithms.

    - average over sequentially constructed predictions.
    - solution path similar to Lasso.
    - learner = linear model. Averaging reduces bias.

c. Random forest/bagging type algorithms.

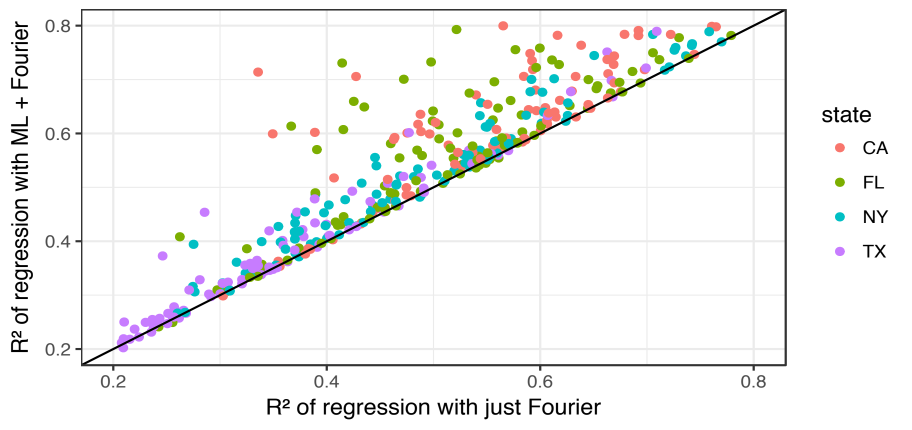
    - average over predictions from randomly chosen predictors.
    - learner = regression tree. Averaging reduces variance.

Adjusted Data: Some Observations

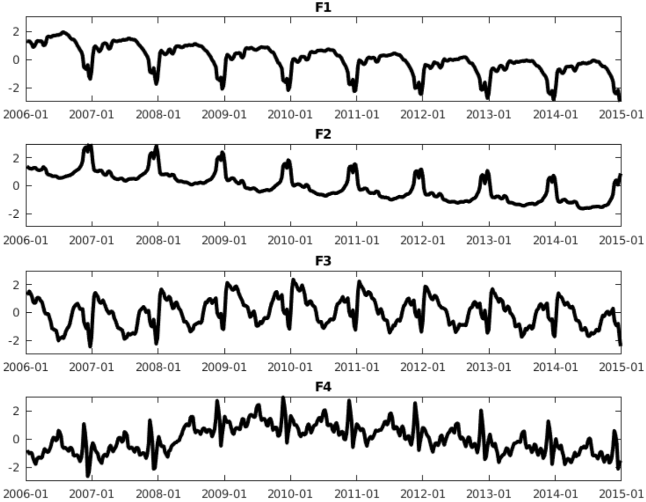
- Largest change: shares in 'insectisides', 'ice', 'stationary', 'candies', 'fragrance'.
- Value added of Step 2: Evaluate R^2 of $y_{gct} = \alpha_{p0} + \alpha_{g1} \cdot \widehat{d}_{gct} + \alpha_{g2} \cdot \widehat{q}_{gct} + u_{gct}$.

	Average of R^2					
method	mean	median	max	q75	q25	min
Fourier	0.49	0.49	0.91	0.60	0.37	0.14
OLS	0.50	0.50	0.91	0.61	0.37	0.14
Lasso	0.52	0.53	0.91	0.65	0.39	0.14
Random Forest	0.54	0.55	0.91	0.68	0.39	0.14

Incremental Predictive Power of Random Forests

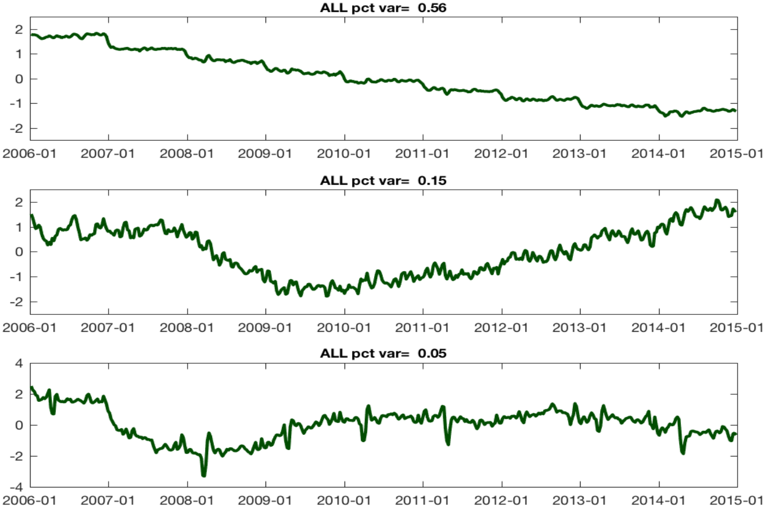


Factors Estimated from Raw Shares: All States



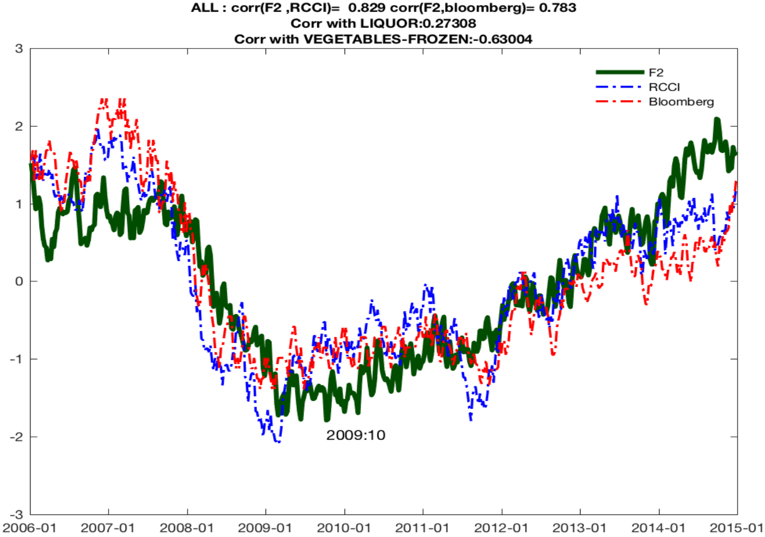
Strong and heterogeneous seasonal effects! Where is the cycle?

Factors Estimated from Adjusted Shares: ALL States

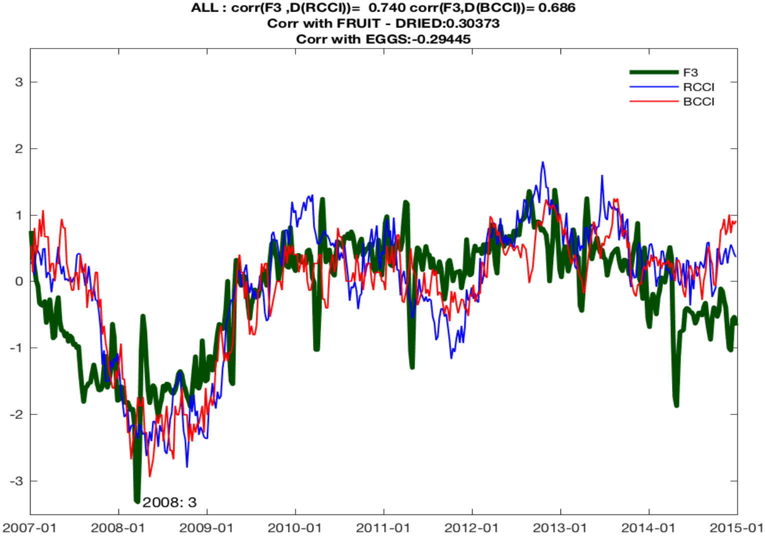


Trend, Level, Slope Factors.

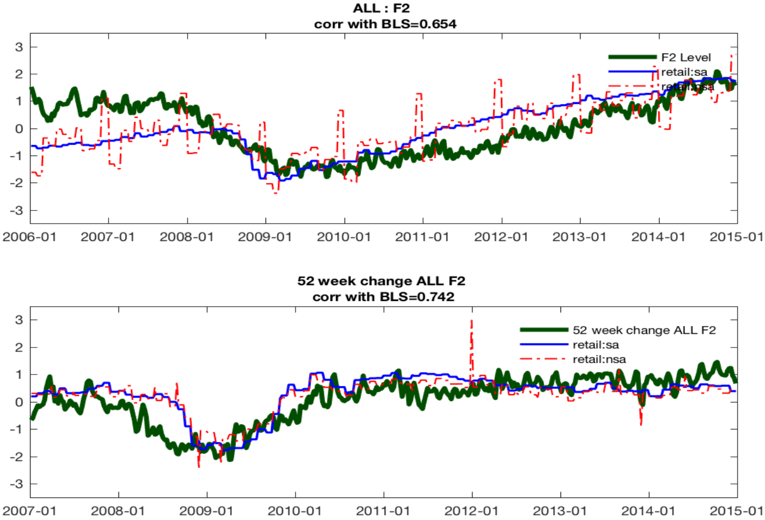
The Level Factor



The Slope Factor



Comparison with the BLS Monthly data

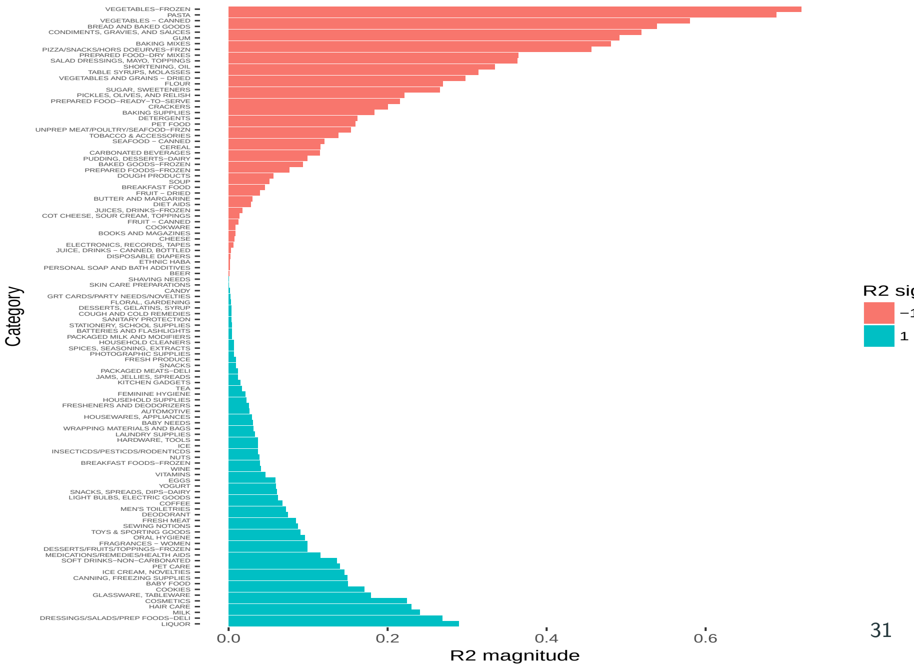


Recession Sensitivity

- Product and regional level information at weekly frequency make the data unique.
- Which product groups are recession-proof? (group,week) panel regression:

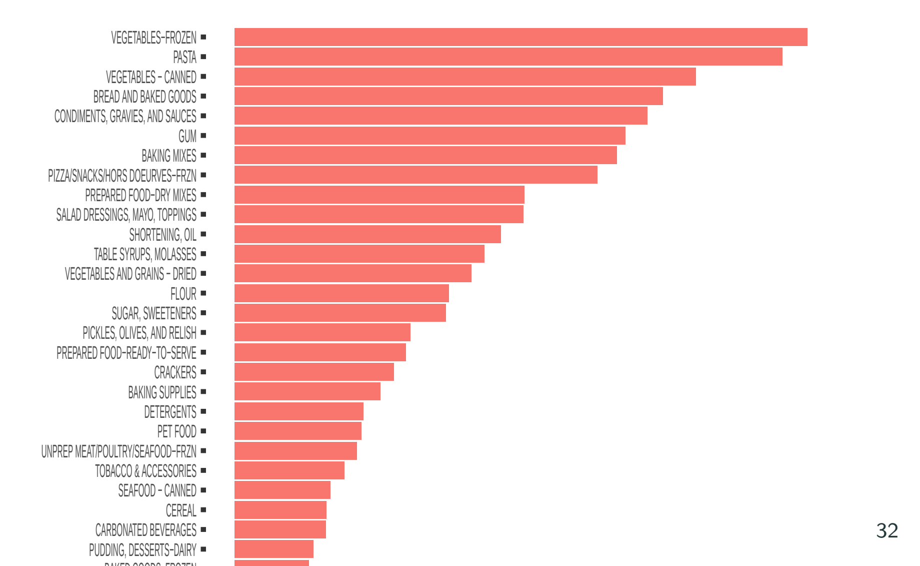
$$share_{gt} = a_0 + a_1 \widehat{F}_{2,RF,t} + error_{gt}$$

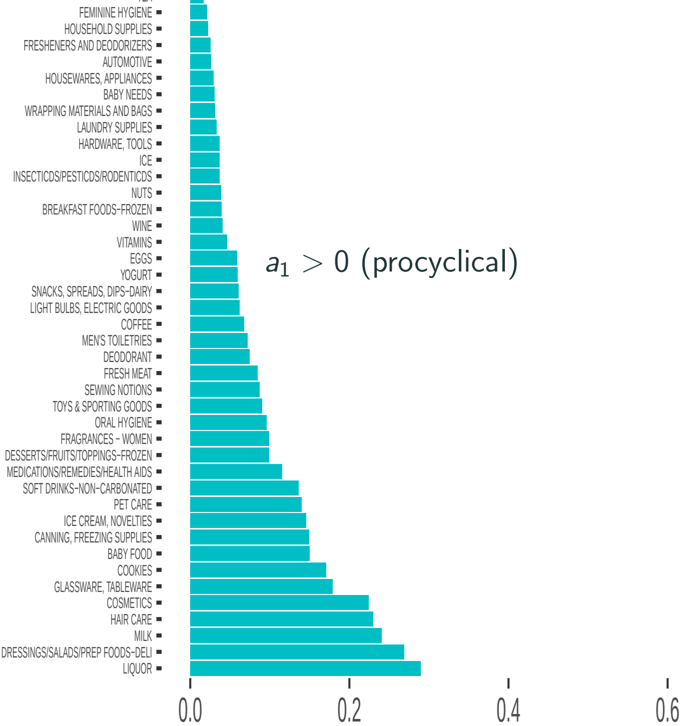
$R^2 \approx 0$: beer, cough/cold remedies, disp. diapers, batteries



Distribution of R^2

$a_1 < 0$ (countercyclical)





$a_1 > 0$ (procyclical)

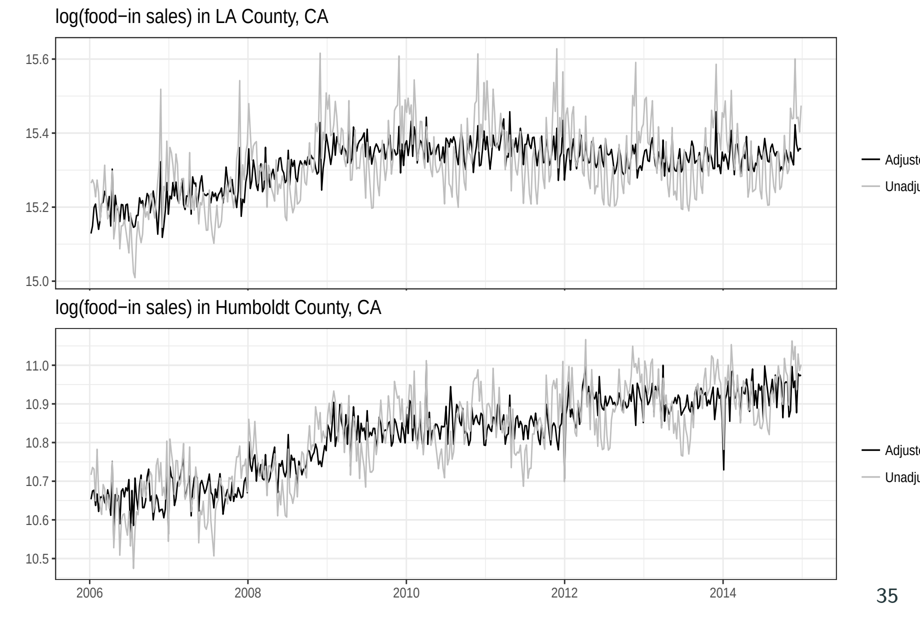
County-level heterogeneity

- Which counties are more exposed to aggregate risk?
- food-in: frozen and canned vegetables, pasta, bread, condiments and sauces.
- luxury: liquor, prepared food, milk, hair care, cosmetics.
- (week, county) panel regressions:

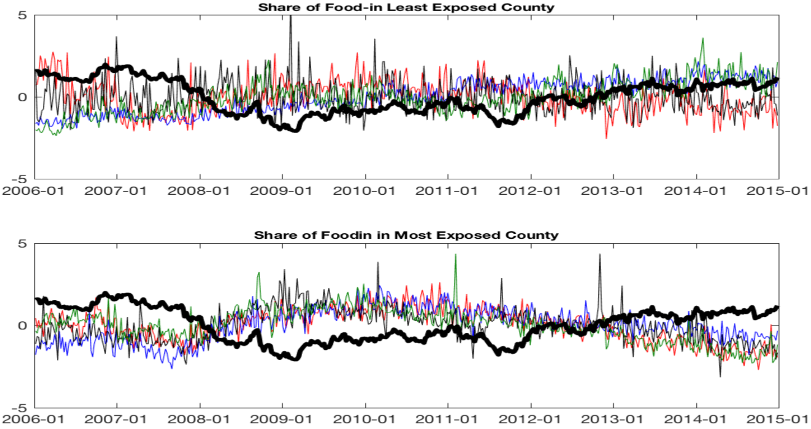
$$\begin{aligned} \text{food-in}_{ct} &= a_1 + a_2 \widehat{F}_{2,RF,t} + a_3 \widehat{F}_{3,RF,t} + error_{ct} \\ \text{luxury}_{ct} &= a_1 + a_2 \widehat{F}_{2,RF,t} + a_3 \widehat{F}_{3,RF,t} + error_{ct} \end{aligned}$$

use R^2 as measure of exposure to common shocks.

LA and Humboldt Counties



Heterogeneity Across Counties: Food-in



Share of Food-in Least Exposed County

Share of Foodin in Most Exposed County

Recession Sensitivity: county level

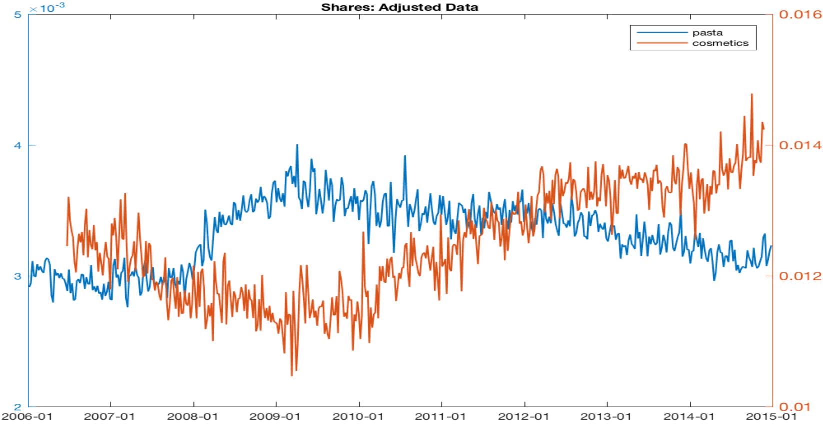
county	CA	FL	NY	TX
	Most sensitive			
1	santa barbara	miami-dade	rockland	montgomery
2	los angeles	broward	nassau	dallas
3	orange	organge	kings	travis
	Least sensitive			
1	sutter	sarasota	seneca	callahan
2	kings	hamilton	lewis	willacy
3	humbodt	lafayette	broome	bexar

urban and densely populated counties are more exposed to aggregate shocks.

Findings so far

- Methodology: A two-step approach that uses algorithms to detect the common information across counties and groups.
    - other applications: predict county level employment using information at neighboring counties.
    - Tweedie-Efron likelihood approach: Koenker-Gu, LMS.
- Empirical Findings
    - Seasonal variations dominate the data, but adjusted data have two cyclical factors: level, and slope factors.
    - Recession-proof analysis of product groups and counties.
        - Food-in group and urban counties most sensitive to $\widehat{F}_2$

Shares of Pasta and Cosmetics: Adjusted Data



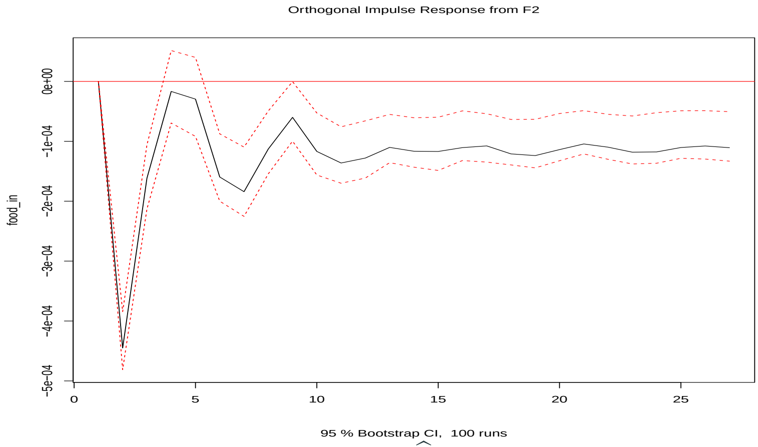
Predictability of County Level “food-in”

Food-in= frozen and canned vegetables, pasta, bread, sauces.

	6 lags				
	(1)	(2)	(3)	(4)	(5)
hline state	food-in	$(1) + \text{rcci}$	$(2) + \widehat{F}_2$	$(2) + \widehat{F}_3$	$(2) + \widehat{F}_2 + \widehat{F}$
CA	0.742	0.744	0.885	0.781	0.901
FL	0.562	0.567	0.666	0.587	0.670
NY	0.497	0.511	0.846	0.643	0.853
TX	0.638	0.641	0.805	0.655	0.854

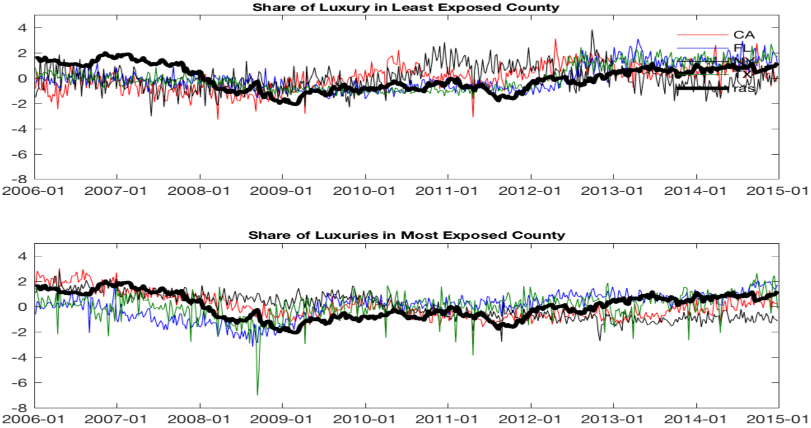
$\widehat{F}_2, \widehat{F}_3$ predict share of food-in, but not RCCI.

Response of food-in to $\widehat{F}_2$ in VAR



- FEVD at h=12: 44% own lags, 48% $\widehat{F}_{2,RF}$.
- RCCI accounts for little of variations in food-in even in bivariate VAR.

Heterogeneity Across Counties: Luxury



Share of Luxury in Least Exposed County

Share of Luxuries in Most Exposed County