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

July 4, 2018

Rishab Guha Harvard University
Serena Ng Columbia University and NBER

# Outline

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

# Macroeconomic Data

- Conventional data
  - Intentionally collected
    - Cleaned, adjusted, and checked
    - Compact and regular.
    - Mostly metrical (not qualitative) variables.
- New Data
  - Often unintentionally generated as 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):
- $N_p \approx$1000 products from $N_g \approx$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. Why?
    - Ng (2017): price and quantity at (store,upc) level.

# This Paper

- Questions:
  - What can the data tell us about small purchases especially around 2008?
  - Is there value in monitoring this weekly data?
- Level of Aggregation: sales of product-groups (or groups).

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

for $s$= CA, FL, NY, TX.

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 |

- Need a framework to summarize broad-based information.

# Demand Systems

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

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

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

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

# Flexible Demand Systems

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

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

  - $p$ and income are common across goods.
  - $P$ imposes cross-equation restrictions. Use proxy simplifies.

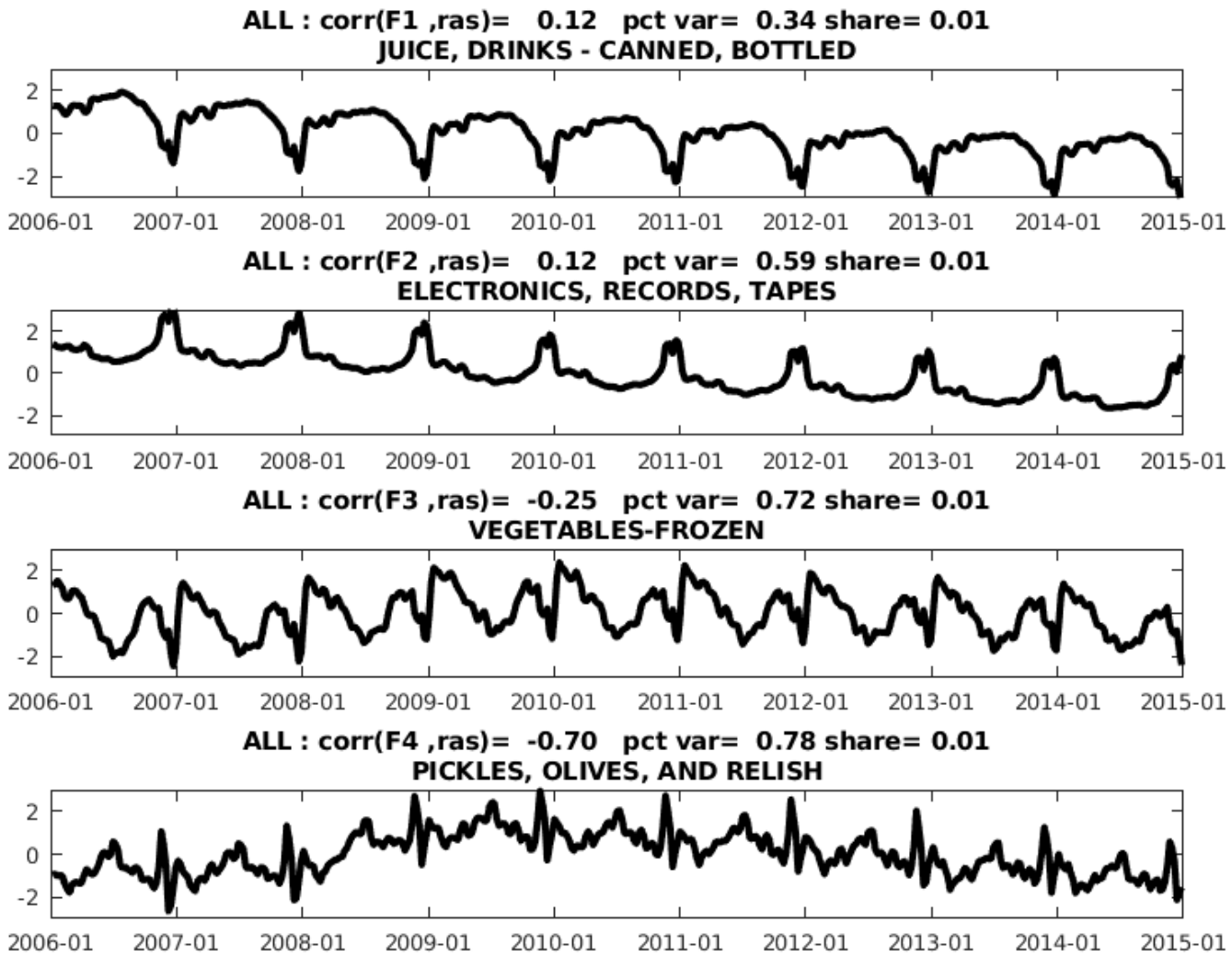
$$\text{share}_g = \lambda_{0g} + \sum \lambda_{jg} \log p_g + \beta_g \log(\text{income}/P^*) + e_g.$$

    In AIDS, $\log P^* = \sum_j w_j \log p_j$ is Stone's price index.
  - $e_g$: measurement error or omitted time varying preferences.

# 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 $s$.
  - share $= F\Lambda'$ : budget shares have a factor structure.
  - estimate latent $F$ and $\Lambda$ without using price/income data
  - non-parametric in economic and econometric sense.
  - rank of demand system in Nielsen data $> 5$. Why?

# Factors Estimated from Data NSA: All




Strong and heterogeneous seasonal effects! Where is the cycle?

# Dealing with Seasonality

- Spending is concentrated in the last 6 weeks of year.
  - in demand system, income=spending=$\sum_g p_g q_g$ 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 $N_{\text{years}} = 9$.

# Existing Methods

- Assume $T$ large, number of seasonal observations large.
- Univariate procedures: X13, SEATS/TRAMO.
    - parametric, require tuning parameters to each series.
- Need an automated and scalable seasonal adjustment method.

# Problem with Existing Approaches?

- Let $y_{ct} = \log \text{sales}_{gct}$.
- Consider $\widehat{seas}_{ct}$ obtained by some univariate scheme.
- Define $\theta_c = \text{corr}(\text{seas}_{ct}, \widehat{\text{seas}}_{ct})$.

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

Suppose for each $c$, $\theta_c \approx 1$.

# Common Seasonality ⇒ Common Adjustment Error

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

- Aggregation:

$$
\begin{aligned}
\widehat{\text{cycle}}_t &= \frac{1}{N_c}\sum_{c=1}^{N_c} \text{cycle}_{ct} + \frac{1}{N_c}\sum_{c=1}^{N_c}(1-\theta_c)\text{seas}_{ct} \\
&= \overline{\text{cycle}}_t + \frac{1}{N_c}\sum_{c=1}^{N_c}(1-\theta_c)\left(\lambda_c^{seas\prime} F_t^{seas} + e_{ct}^{seas}\right) \\
&= \overline{\text{cycle}}_t + \overline{(1-\theta_c)\lambda_c^{seas\prime}} F_t^{seas} + o_p(1).
\end{aligned}
$$

- unless filter is ideal $\forall c$, or $\lambda_c^{seas} = 1$, bottom up leaves $F_t^{seas}$ in estimate of cycle. Ng (2017).

# Implications

- Lots of spikes from holiday sales at dates that are not fixed across years.
- Smooth functions are not good at picking up spikes.
- Even if we have many years of data, parametric time series adjustements will not likely be perfect.
- We expect common seasonality in aggregate adjusted data.

# A New Approach

- Given goal of recovering broad-based information, we remove seasonality year by year using a panel approach, rather than the variable by variable conventional time series approach.
- 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:

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

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

- Key: seasonal components $d_{gct}$ and $q_{gct}$ are predictable.
- We 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 $\widehat{q_{gct} + u_{gct}}$

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

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 $(g, s)$: want to predict $q_{gct}$ for all weeks $t \in \tau$.
- Control for multidimensional sesaonal heterogeneity using lots of dummy predictors using a flexible seasonality function.
  - Intuition from fixed effect: 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: holidays, sports events.

  ii social-economic indicators at county level.

  iii weather and location from NOAA.

  iv interaction of (i) and (ii).

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

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

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

## 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.

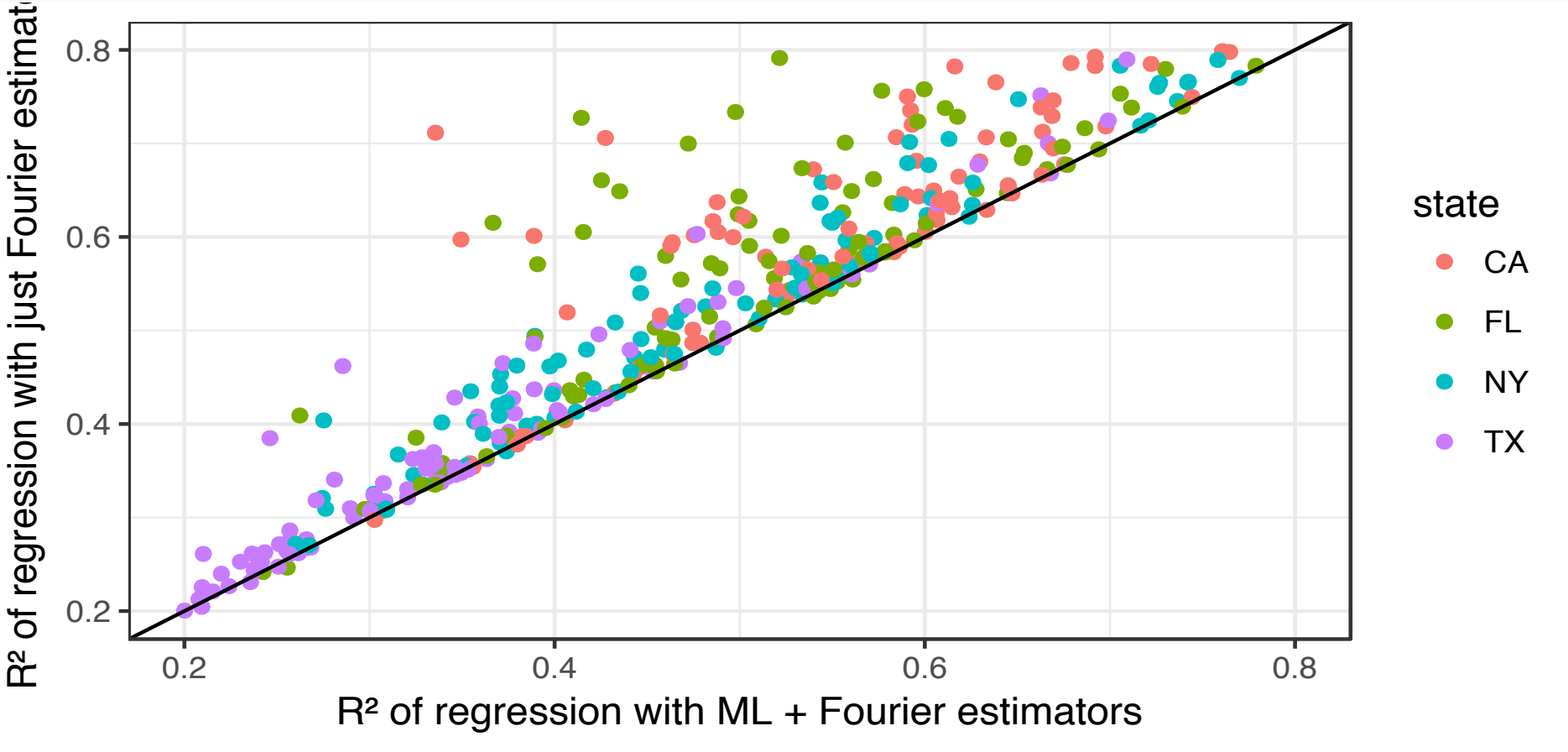
- average over predictions from randomly chosen predictors.
- learner = regression tree. Averaging reduces variance.

# Adjusted Data: Some Observations

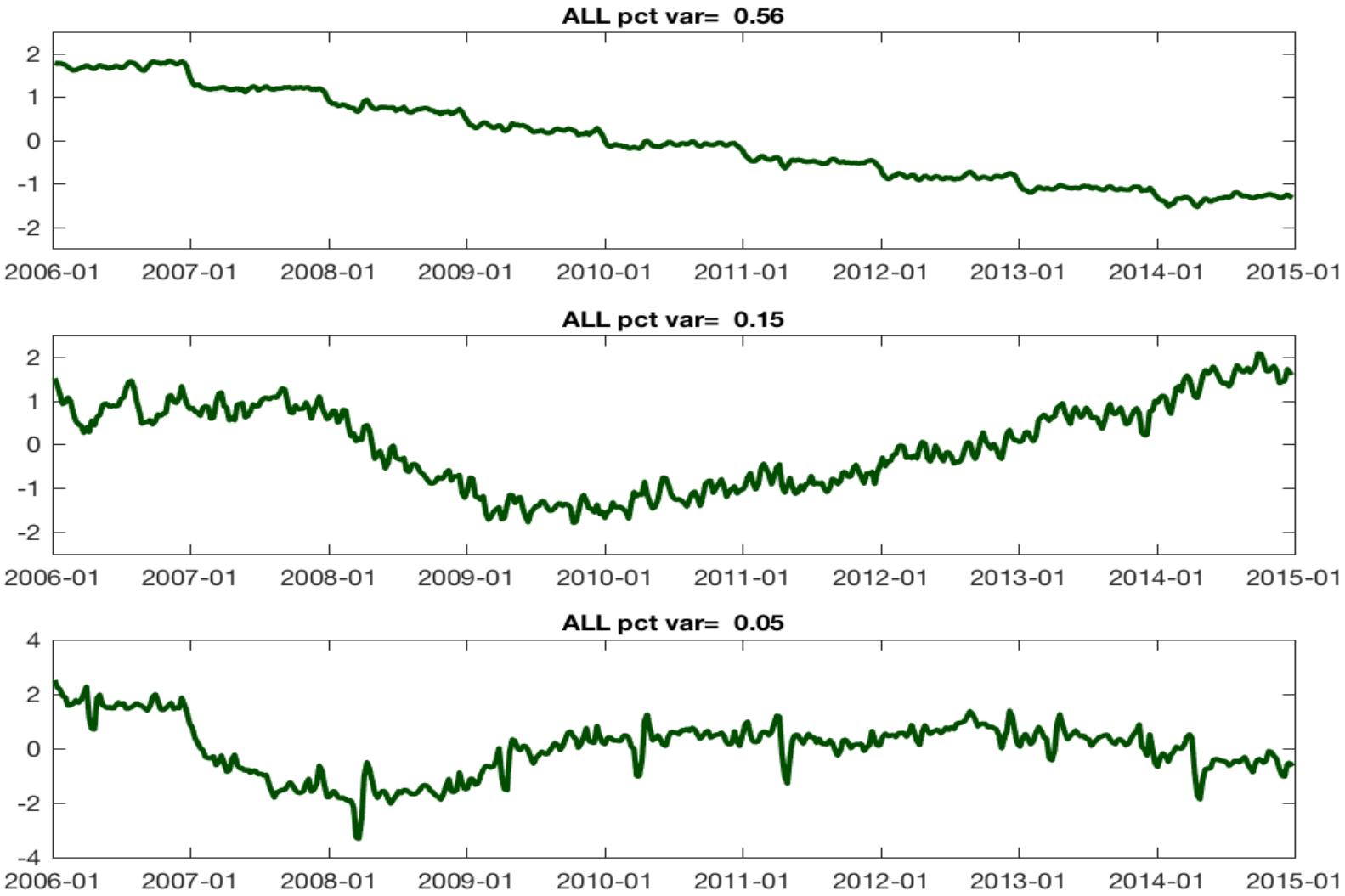
- Largest change: shares in 'insectisides', 'ice', 'stationary', 'candies', 'fragrance'.
- How much predictability does Step 2 (ML) add over Step 1 (Fourier)? Compare R2 from regressions of $y_{gct}$ on $x_{gct}$.

| | 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 |

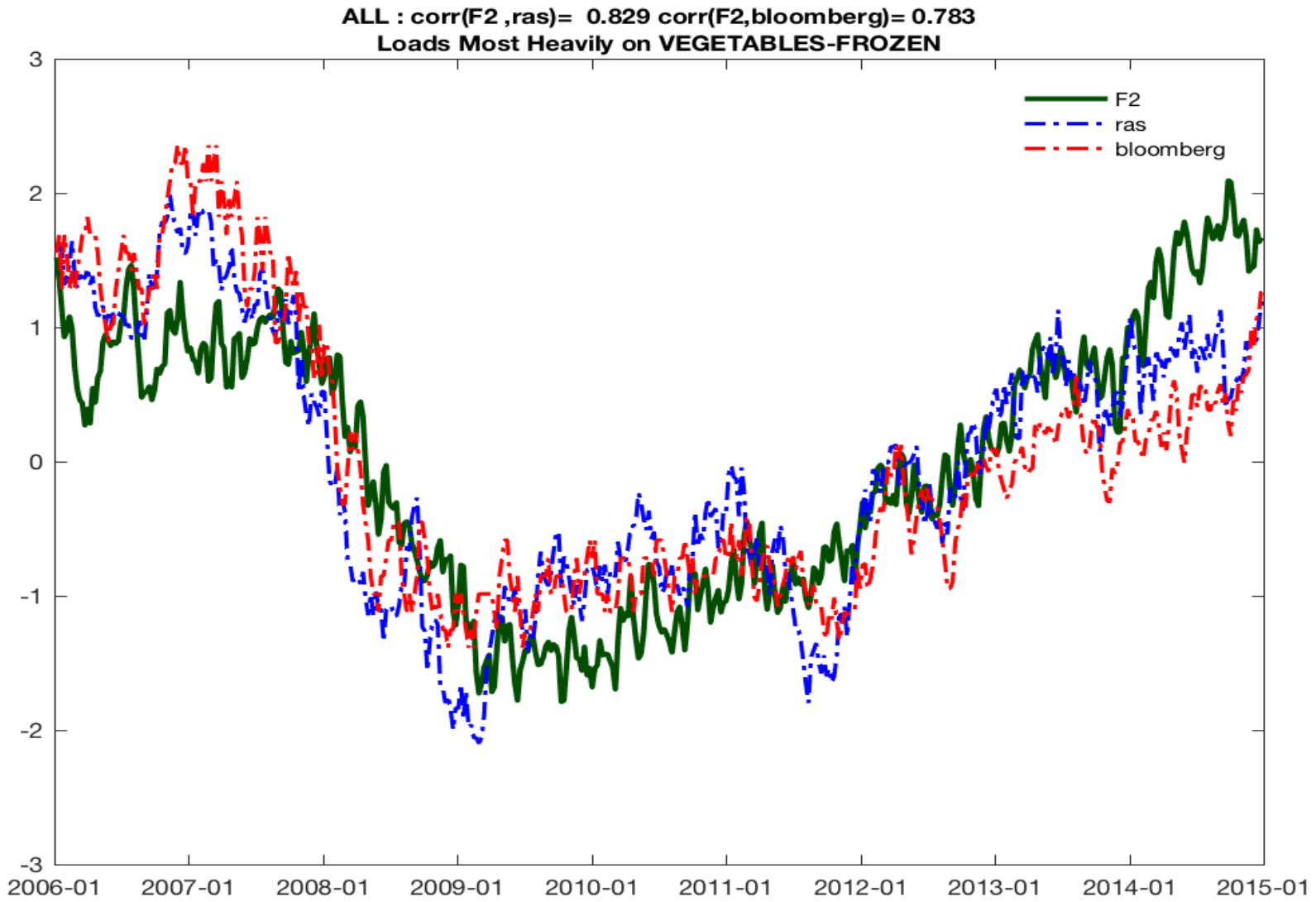
# Incremental Predictive Power of Random Forests



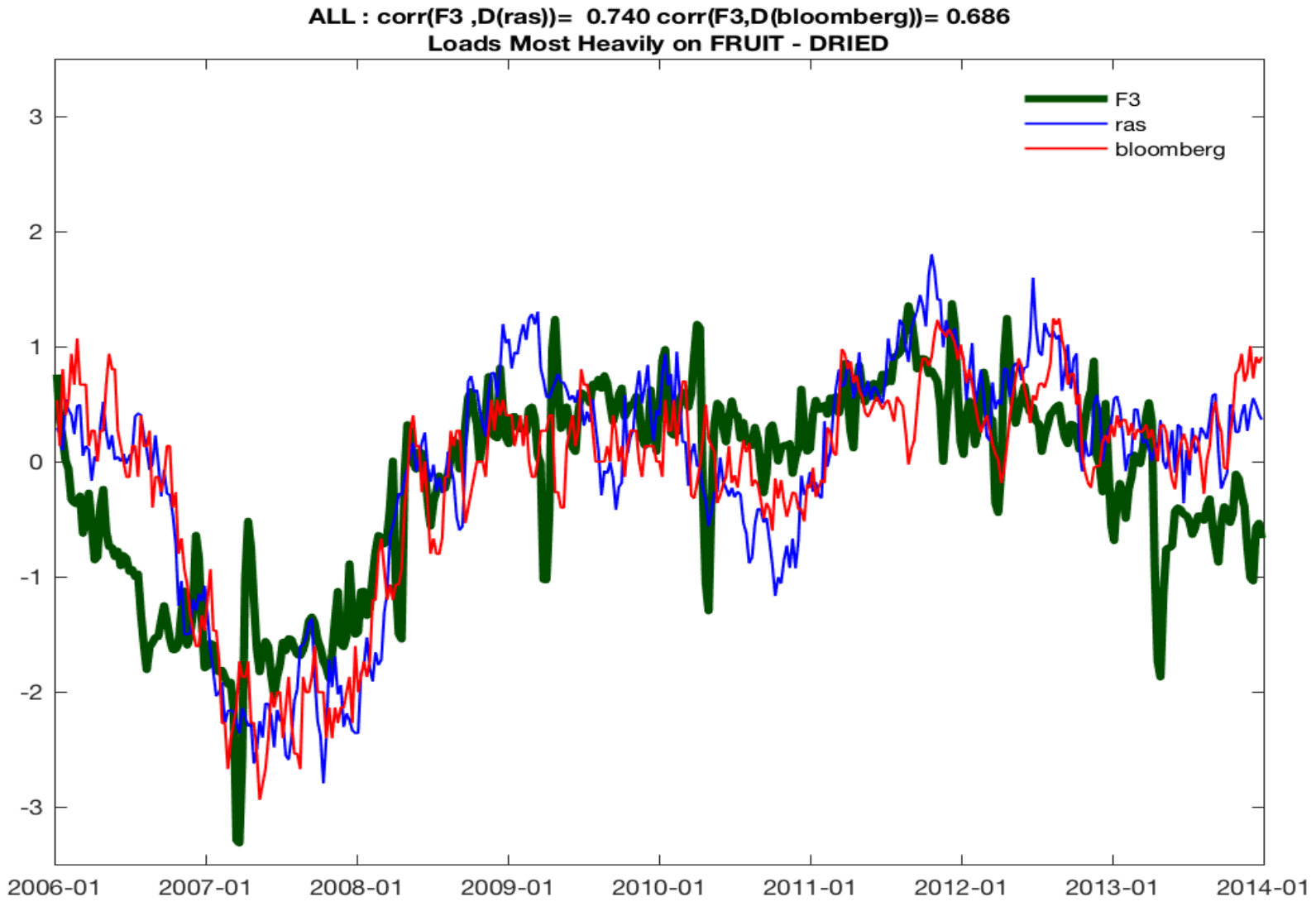
# Factors Estimated from Adjusted Shares: ALL States



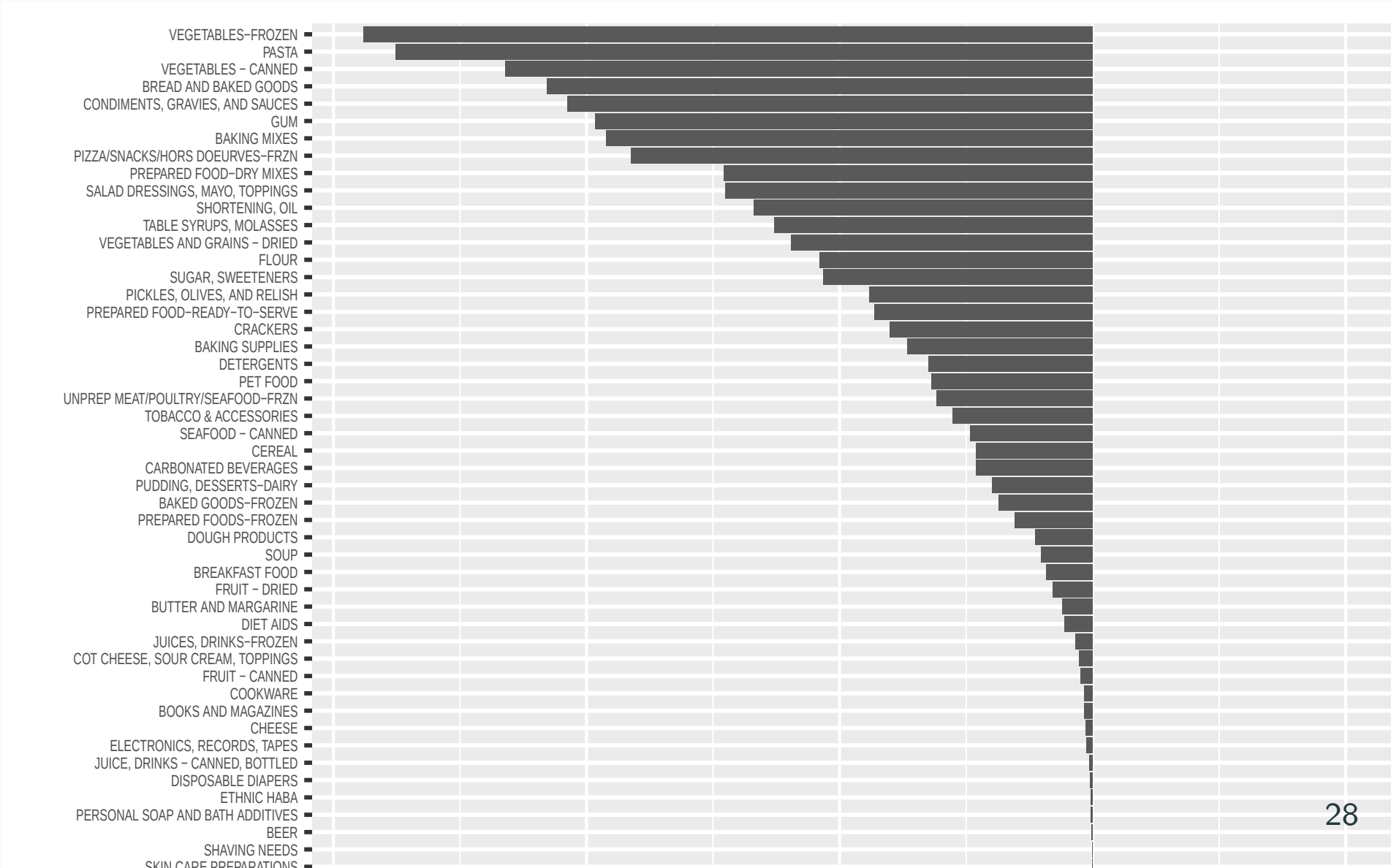
# The Cyclical Factor

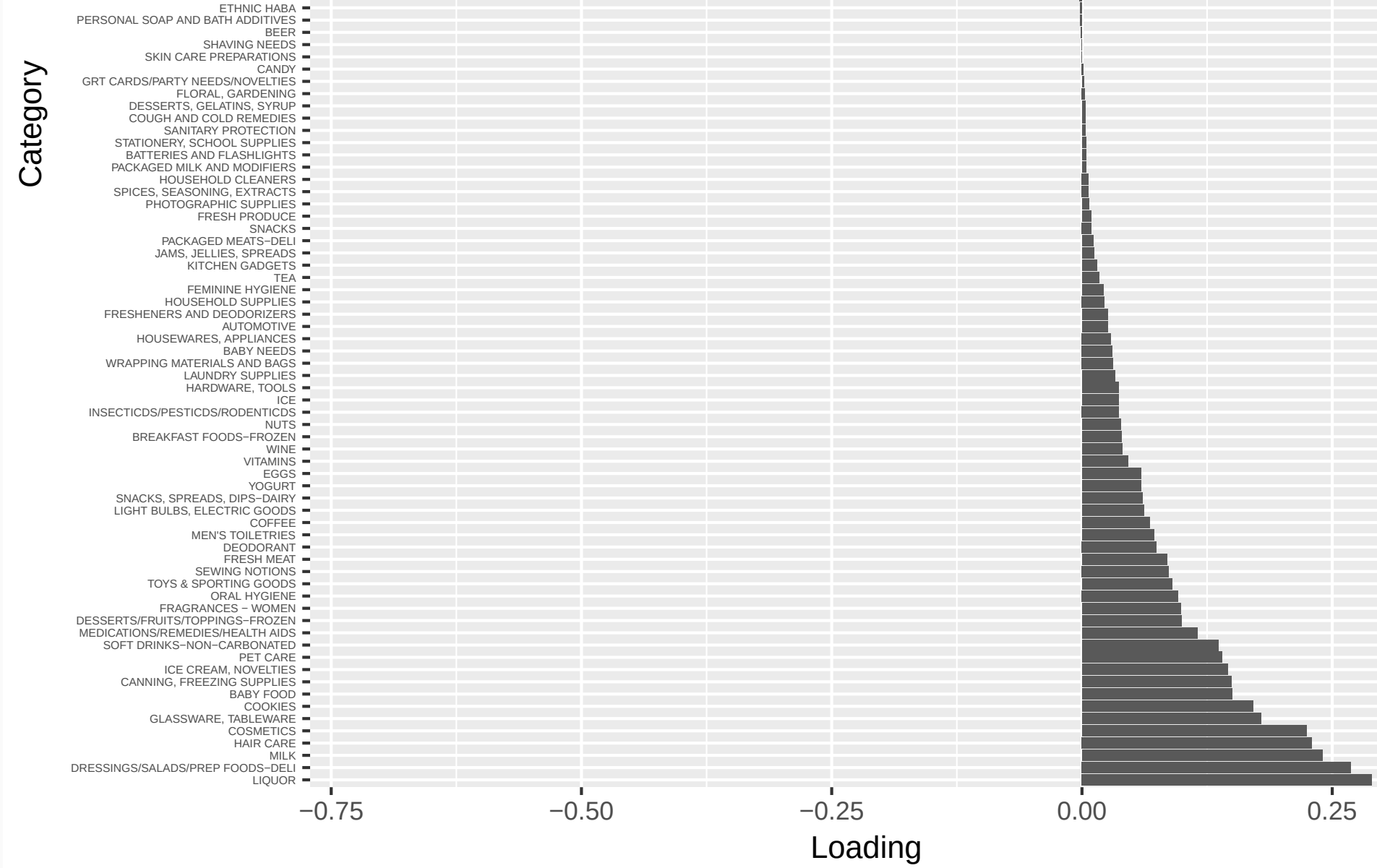


# The Direction Factor



# Distribution of Estimated Loadings




Category
ETHNIC HABA
PERSONAL SOAP AND BATH ADDITIVES
BEER
SHAVING NEEDS
SKIN CARE PREPARATIONS
CANDY
GRT CARDS/PARTY NEEDS/NOVELTIES
FLORAL, GARDENING
DESSERTS, GELATINS, SYRUP
COUGH AND COLD REMEDIES
SANITARY PROTECTION
STATIONERY, SCHOOL SUPPLIES
BATTERIES AND FLASHLIGHTS
PACKAGED MILK AND MODIFIERS
HOUSEHOLD CLEANERS
SPICES, SEASONING, EXTRACTS
PHOTOGRAPHIC SUPPLIES
FRESH PRODUCE
SNACKS
PACKAGED MEATS−DELI
JAMS, JELLIES, SPREADS
KITCHEN GADGETS
TEA
FEMININE HYGIENE
HOUSEHOLD SUPPLIES
FRESHENERS AND DEODORIZERS
AUTOMOTIVE
HOUSEWARES, APPLIANCES
BABY NEEDS
WRAPPING MATERIALS AND BAGS
LAUNDRY SUPPLIES
HARDWARE, TOOLS
ICE
INSECTICDS/PESTICDS/RODENTICDS
NUTS
BREAKFAST FOODS−FROZEN
WINE
VITAMINS
EGGS
YOGURT
SNACKS, SPREADS, DIPS−DAIRY
LIGHT BULBS, ELECTRIC GOODS
COFFEE
MEN'S TOILETRIES
DEODORANT
FRESH MEAT
SEWING NOTIONS
TOYS & SPORTING GOODS
ORAL HYGIENE
FRAGRANCES − WOMEN
DESSERTS/FRUITS/TOPPINGS−FROZEN
MEDICATIONS/REMEDIES/HEALTH AIDS
SOFT DRINKS−NON−CARBONATED
PET CARE
ICE CREAM, NOVELTIES
CANNING, FREEZING SUPPLIES
BABY FOOD
COOKIES
GLASSWARE, TABLEWARE
COSMETICS
HAIR CARE
MILK
DRESSINGS/SALADS/PREP FOODS−DELI
LIQUOR
−0.75
−0.50
−0.25
0.00
0.25
Loading

# Predictability of County Level "food-in"

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

| | 6 lags | | | | |
|---|---|---|---|---|---|
| hline state | (1) food-in | (2) (1) + rcci | (3) (2) + $\widehat{F}_2$ | (4) (2) + $\widehat{F}_3$ | (5) (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 |

$\widehat{F}_2, \widehat{F}_3$ Granger cause share of food-in.

# Conclusions

More results: To be completed.