A Machine Learning Analysis of Seasonal and Cyclical Sales in Weekly Scanner Data

Rishab Guha*
Harvard University

Serena Ng†
Columbia University and NBER

July 9, 2018
Preliminary and Incomplete
Comments Welcome

Abstract

This paper provides methods to summarize information in weekly sales of 108 groups collected at the county level between 2006 and 2014. The data display multi-dimensional heterogeneity with seasonal effects that are convolved with cyclical variations. The short span and large volume of data make it impractical to seasonally adjust the data by structural time series modeling, and in any event, these methods are not suited for weekly data because of its quasi-periodic nature. We suggest a panel data approach that pools information across counties and over years. We treat seasonal adjustment as a prediction problem and use machine learning methods to find the best model and predictors. The demand system estimated from seasonally adjusted sales has a rank of three. There is a distinctly cyclical factor that moves closely with consumer confidence and another factor that matches up well with the change in consumer confidence. The sensitivity to this factor distinguishes the pro-cyclical from the counter-cyclical groups. Consumption shares at the county level have predictive power for the one-month ahead county level unemployment. Thus once the seasonal variations are removed, the data is a rare source of timely information about local economic conditions.

Keywords: Seasonal adjustment, Random forest, fixed effects, demand system.

*Littauer Center, Harvard University, 1805 Cambridge Street, Cambridge, MA 02137. Email: rishab_guha@g.harvard.edu

†Department of Economics, Columbia University, 420 W. 118 St. MC 3308, New York, NY 10027. Email: serena.ng@columbia.edu

Finally support from the National Science Foundation (SES-1558623) is gratefully acknowledged.

1 Introduction

A mind-boggling amount of data is now available for economists to analyze. New, post-Google, data sources can be differentiated from the data sources traditionally used by economists in at least two important ways. They tend not to be provided by government agencies and have what data scientists refer to as the "three V" characteristics: volume, variety, and velocity. Econometricians may think of them as short panels of big, often unbalanced, high-frequency, highly heterogeneous data. Such granular data can potentially allow for a better understanding of how micro-level behavior affect aggregate outcomes compared to conventional data sources. However, as discussed in Ng (2017), a full analysis of the data comes with unique challenges. A case in point is the weekly Nielsen Retail Scanner data collected since 2006, and which is now over five terabytes in size.

The Nielsen data have three features of interest. First, the data are real-time sales and unit price recorded at the store level. Such transactions data are distinctly different from published price indices that necessarily depend on modeling assumptions. They are not subject to revisions once a transaction is completed. Second, the data are available for the major metropolitan areas on a weekly basis. It thus provides spatial and temporal information that is absent from the monthly retail trade data released by the U.S. Census. Third, the 2006-2014 sample includes a big recession, a memorable hurricane, and a government shutdown. A lot about consumer behavior can potentially be learned from the data. But macroeconomic analyses using this data thus far have mostly aggregated the data to a monthly or quarterly frequency, thereby surrendering the weekly information that makes the data unique. With just a peek of the data, one sees why:- the data exhibit strong seasonality that is highly heterogeneous in the product and spatial dimensions. To use this data for business cycle analysis, one needs to deconvolve the seasonal and cyclical variations, and there are few known methods for seasonally adjusting weekly data, let alone for a massive number of series.

This paper develops methods to extract information about the state of the local and aggregate economies from the disaggregated data. We isolate the common variations by studying budget shares as a demand system. In a demand system, Engel curves are spanned by functions of prices and income that are common across product groups. For each state, we have data for N_g=108 product groups over T =469 weeks at the county level. As N_g and T are large, we use methods developed in large dimensional factor analysis to estimate the latent functions of prices and income without using data on either prices or income. The common factors estimated from the raw data for California (CA), Florida (FL), New York

(NY), and Texas (TX) indicate that the leading factors are overwhelmingly dominated by seasonal variations. To learn about the cyclicality of sales necessitates that we find a way to remove the seasonal variations.

We develop a panel data approach that can seasonally adjust a large panel of data with a short span. Our point of departure is that while county location, product features, and time effects contribute to heterogeneous seasonal variations, much of these variations are highly predictable *ex-ante*. Hence we treat seasonal adjustment as a prediction problem. To find the prediction model of unknown functional form in the face of a large set of potentially relevant predictors, we use machine learning methods to perform estimation and variable selection. This by-passes the need to specify a single data generating process, which is a difficult task when the data are so highly heterogeneous. The method differs fundamentally from existing univariate regressions or filtering methods which rely on long spans of the data to remove strictly-periodic seasonality.

Our analysis leads to four conclusions. First, most of the seasonal variations in the data are predictable. However, even though the data are primarily based on sales at grocery and mass merchandise stores, there is surprisingly clear evidence of cyclical spending patterns in the seasonally adjusted data. Second, the cyclical component in the seasonally adjusted data moves closely with measures of consumer sentiment and confidence, indicating that the actions and 'feelings' of consumers are aligned. Third, the loadings on the cyclical factor imply a 'distribution' of recession sensitivity across product groups. The budget share of products such as pasta and frozen vegetables tend to be particularly sensitive to the cyclical factor. Fourth, the seasonally-adjusted budget-shares at the county level are found to have predictive power for the one-month ahead change in that county's unemployment rate especially during the Great Recession. The results suggest that the data can be a rare source of high frequency aggregate and local economic conditions once the seasonal variations are controlled for.

2 The Data

The Nielsen Retail Scanner data are collected by the Nielsen marketing group and managed by the Kilts Center for Marketing at the University of Chicago. The data have over 1000 products belonging to over 115 product groups (e.g., beer, wine, eggs) that can in turn be organized into 10 categories: dry groceries, frozen, dairy, deli meat, fresh food, non-food, alcoholic beverage, general merchandise, health and beauty. The data are heavily weighted towards groceries and mass-merchandise goods with limited coverage of consumer durables.

Specifically, the products cover over 3 million UPCs collected from over 35,000 participating stores in 55 MSAs across the U.S. We have information about the location of the retailer (but not the name), the units sold and the price, from which we can compute the value of sales. Following Nielsen's documentation[1] the week's total dollar sales is

$$\text{sales} = \frac{\text{price}}{\text{prmult}} \times \text{units}.$$

The *movement files* of the database provide data for UNITS (the number of units sold), PRICE (the volume weighted average price of the product for the week), PRMULT (a price multiplier to indicate deals such as 3 for one dollar).

We analyze the total sales of products within a group. Hereafter, we will simply refer to them as 'groups' instead of 'product groups'. The primary focus will be on the four most populated states in the U.S: California (CA), Florida (FL), New York (NY), and Texas (TX). The sales data are constructed as follows. For each state, s, we first compile the list of stores that report a sale in at least one of the 115 groups in each of the 469 weeks in the sample from January 2006 to December 2014. Restricting attention to groups with data in every week reduces the number of groups from 115 to $N_g = 108$. This ensures a balanced panel of stores over our sample period. The variables will be indexed as follows:

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

Group by group, we construct a measure of weekly total sales at the county level by aggregating over all stores for each county in the compiled list. The variable of interest is $\log(\text{SALES}_{gct})$, the log sales of group g in county $c(s)$ in week t[2]. Since a county is state-specific, the state index s will be suppressed when the context is clear. We also construct sales by aggregate sales over the four sates. This is labeled 'ALL' below. At each week t, the budget share of an arbitrary group g is

$$\text{share}^s_{gt} = \frac{\sum_{\{c:\mathbf{s}(c)=s\}} \text{SALES}_{gct}}{\sum_p \sum_{\{c:\mathbf{s}(c)=s\}} \text{SALES}_{gct}} = \frac{\text{sales of group g in state s at week t}}{\text{total sales in state s at week t}}$$

The ten products with largest sales are listed below.

[1] Kilts Nielsen Retail Scanner Dataset Manual 2014, Table 3, Part E.

[2] Throughout the text we use "county" to reference a specific geographic county, e.g., New York County, New York. We let $\mathbf{s}(i)$ denote the state containing county i

Budget Shares: Most Purchased Categories

CA: $N_c = 53$		FL: $N_c = 58$		NY: $N_c = 58$		TX: $N_c = 161$	
0.034	bread	0.044	medications	0.041	medications	0.037	carbon. bev
0.033	beer	0.043	tobacco	0.032	fresh produce	0.037	medications
0.033	juice	0.031	carbon. bev.	0.031	bread	0.034	snacks
0.032	wine	0.029	liquor	0.030	candy	0.029	bread
0.030	fresh produce	0.028	beer	0.028	snacks	0.028	tobacco
0.030	carbon. bev	0.026	juice	0.028	juice	0.026	packaged meat
0.030	snacks	0.026	candy	0.027	tobacco	0.026	candy
0.027	packaged meat	0.024	snacks	0.025	beer	0.025	fresh produce
0.027	salad dressing	0.023	milk	0.024	carbon. bev	0.025	juice
0.026	medication	0.026	bread	0.023	milk	0.025	beer

There products are reasonably similar, with bread, beer, juice, carbonated beverages, medication, and snacks making the list in all four states.

A more systematic analysis of the data requires a framework, and we appeal to demand analysis. A demand system expresses N_g budget shares as functions of prices (denoted $p = (p_1, \ldots, p_{N_g})'$) and income (denoted income). If we denote these functions by F and let Λ be the sensitivity to these functions, then under assumptions of demand theory, budget shares have representation

$$\underbrace{\text{share}}_{N_g \times 1} = \underbrace{\Lambda}_{N_g \times r} \underbrace{F(\log p, \log(\text{income}/P))}_{r \times 1}$$

where P is a price index consistent with the demand system. Under two-stage budgeting, the income equals total expenditure on the N_g groups. Demand theory forms the basis for constructing measures of cost of living.

The functions F functions are common across groups. The rank of a demand system r is the dimension of the space spanned by $F = (F_1, \ldots, F_r)$ where the adding up constraint requires F_1 is a constant.[3] A rank one system occurs when budget shares are independent of the level of income in which case all income elasticities equal one. Rank two demand systems are linear in log-prices but not in log income. Examples include the translog, linear expenditure systems. Many rank two systems belong to the PIGLOG class, see Muellbauer (1975). Quadratic Engel curves can be rank two or rank three. Gorman (1981) shows that exactly aggregable demand systems must have rank no larger than three.

In empirical work, the cross-equation restrictions imposed by P is relaxed by using a proxy P^*. For example, the Almost Ideal Demand System of Deaton and Muellbauer (1980).

[3]For a discussion on the rank of demand systems, see Lewbel (1991).

performs linear estimation of

$$\text{share}_{gt} = \lambda_{0g} + \sum_{j=1}^{N_g} \lambda_{jg} \log p_{gt} + \beta_g(\log(\text{income}_t/P_t^*)) + e_{gt}$$

using data for shares of dimension $T \times N_g$. With cross-sectional analysis, T is number of observations on household level spending. For time series analysis, T is the number of time series observations on aggregate spending. In cases when panel data are available, T is the number of household data stacked over time. In these studies, N_g is small and T is large. Under these assumptions, the typical estimate of rank is between two and four, see Lewbel (2003, 1997); Banks et al. (1997).

Two features make demand estimation using the Nielsen data distinct. First, the term e_g is usually attributed to measurement or specification error such as omission of time variation in preferences. But with scanner data, measurement error is less likely. Second, there is no hope of recovering Λ, F and e without parametric assumptions when N_g is small. But this is possible in our setup because our sample size is large in both the time (weeks) and cross-sectional (group) dimensions. Instead of imposing parametric assumptions and using proxy variables, we treat F and Λ as latent variables. According, we consider the factor representation of budget shares:

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

where F_t is a $r \times 1$ vector of latent factors and Λ_g is the corresponding vector of factor loadings. Importantly, we can consistently estimate the space spanned by the factors using the shares data alone, i.e. without using data on prices and income. Also consistently estimable is the number of factors, which is the rank of the demand system.

Prior to estimation, we take a three week rolling average of the budget shares so as to smooth out the variations due to short-duration promotional sales. Principal components are then estimated from the standardized data.[4] The criterion of Bai and Ng (2017)[5] finds five factors in CA, FL, NY, four factors in TX, and five factors in the pooled panel 'ALL'. In all cases, the first factor explains about one-third of the variations in the data, the first two factors together explain just under sixty percent, while four factors explain around seventy-five percent of the variations in budget shares. These estimates are higher than the ones

[4]Since our data are demeaned, the constant factor is controlled for prior to estimation.

[5]The criterion is defined as $\bar{r}^s = \min_{k=0,\ldots,\text{rmax}} \log\left(1 - \sum_{j=1}^{k}(\sigma_j^s - \gamma)_+^2\right) + k \cdot \text{PENALTY}(N,T)$. where $\text{PENALTY}(N,T) = \frac{(N+T)}{NT}\log(\frac{NT}{N+T})$ and σ_j is the j-largest value i a $T \times N$ panel $Z = \frac{X}{\sqrt{NT}}$, X is a panel of standardized data, in our case, a panel of budget shares. A regularization parameter of $\gamma = 0.05$ is used to penalize common variations due to outliers. The the maximum number of factors is set to 10.

previously obtained from estimation using lower frequency, seasonally adjusted data of much smaller dimensions.

What accounts for the unusually high estimate of the rank? One explanation is that the data used here are not seasonally adjusted. This is important because sales of 'egg-nog', for example, peak around Christmas while sales of stationary and school supplies peak around week 36. The share of spending on 'cough and cold' products peaks around week 8. Flower sales are high around Valentine's Day and Mother's Day. Ice-cream sales are higher in the summer. Beer sales tends to peak around July fourth while wine sales tend to peak around Thanksgiving and Christmas. Notably, these spikes in sales recur every year regardless of location. The factors we estimated from the raw data are reflect pervasive seasonal sales.

The strength of common seasonal variations is best seen from Figure 1, which plots the first four factors from the pooled data ALL. The factors are mutually orthogonal by construction. Hence Figure 1 indicates a multitude of seasonal effects at work. The factors are only identified up to sign, so they are plotted to be pro-cyclical. Though all four factors have spikes that cluster around week 48, the exact week is different over years. Correspondingly, the spectrum of the first few factors estimated for each state has peaks around the seasonal frequencies of $\frac{2\pi j}{52} = .1208j$ for $j = 1, 2, \ldots$ but not exactly at those frequencies. Non-parametric regressions of the factors with sines and cosines as orthogonal basis find over 85% and as much as 95% of the variations in the first two factors can be explained by strictly periodic functions.

Principal components only identify combinations of dominant variations in the data and is blind to the source. That the first few factors are strongly seasonal and not cyclical indicates the relative strength of the two types of variations. To give some interpretation to the seasonal factors, we explore the contribution of the factors to variations in the shares and find that in s=(NY, TX, and ALL), $\hat{F}_1^s$ explains over 0.9 of the variations of PACKAGED-MILK. As egg-nog is in the group, $\hat{F}_1^s$ is arguably a *Christmas* factor. Other products also exhibit Christmas effects. In CA for example, the share of juice takes a big dip in week 51, while in FL, the share of hair-care products bottoms around week 51. The presence of seasonal variations in the factors corresponding to the smaller eigenvalues is clear. Nonetheless, Figure 1 suggests cyclical movements especially in factors three and four. To see if the data are useful for macroeconomic analysis, we need a way to deconvolve the variations at the seasonal and the non-seasonal frequencies. This is the focus of the next section.

3 Seasonal Adjustment as a Prediction Problem

Consumer theory suggests it is desirable to smooth consumption over time. But in reality, spending is concentrated around holidays, special events, and towards the last six weeks of the year. Seasonality has implications for demand analysis because log-income in estimation is replaced by log total expenditure on a given basket of groups considered. If spending exhibits seasonality, so will functions driven by it. A more appropriate representation of budget shares when the data are not seasonally adjusted is

$$\text{share}_{gt} = \Lambda_g^{c\prime} F_t^c + \Lambda_g^{seas'} F_t^{seas} + e_{gt} = \Lambda_g' F + e_{gt}$$

where the superscript 'c' denotes cyclical, and 'seas' denotes seasonal. For applications such as forecasting weekly advertising revenue, the seasonal effect is of direct interest and needs to be carefully modeled. But for business cycle analysis, the seasonal component of demand must be removed.

A variety of univariate methods are available to de-seasonalize monthly and quarterly data. Examples include X13ARIMA-SEATS and TRAMO-SEATS. However, these methods do not seem appropriate for data with the three-V features. For one thing, although we have 469 weekly observations, there are only nine years of data, meaning that we only have nine data points from a seasonal perspective. Time series seasonal adjustment methods typically assume that we have a large number of observations for a given periodicity of interest.

The fact that we have weekly data poses a unique challenge because weekly variations are not exactly periodic. For example, Thanksgiving and Christmas do not always fall on the same numbered week of the year, a consequence of the fact that we are on a Gregorian calendar. Easter it is not even always in the same month. In our sample, Easter was as late as April 24 in 2011, and as early as March 23 in 2008. We cannot 'difference away' the seasonal effects like we could with monthly and quarterly data. Even if we manipulate the calendar to do day-of-year differencing, we would not be able to capture the effects of events that do not take place on exactly the same day of the year. Further complicating the problem is that these events have differential impact depending on the product in question and location of the sale.

The literature for adjusting weekly data is quite sparse. Notable exceptions are the state-space analysis of Harvey et al. (1997), and the non-parametric approach of Pierce et al. (1984) which approximates seasonal component by trigonometric basis functions. Cleveland et al. (2014) generalizes the approach to control for weekly effects, holiday effects and outliers using locally-weighted regressions. While Cleveland et al. (2014) shows that the method is

effective in seasonally adjusting unemployment income claims and steel production data, our scanner data have several features that are distinct from the data used in their illustrations.

There are also features specific to our weekly sales data that require special attention. First, our data tend to be 'spiky'. For many groups, the spikes only occur once per year usually around Black Friday. For other series, the spikes can be observed a few times a year and can be attributed to to temporary sales. Spikes are problematic for econometric modeling as they tend not to be well approximated by non-parametric regressions that are smooth by design. Second, some series do not repeat over the course of the year. Instead, they repeat in reference to a date t's position within the *month*. As an example, sales increases around food stamp distributions, or end-of-month price changes. Strictly periodic functions using position within the year as basis function is too restrictive for these variations.

A third characteristic of our data its volume. We have not one, but a large number of series that need to be adjusted. Because of heterogeneity, it would be hard if not impossible to write down a parametric model that would suit all series. Structural time series modeling requires tuning parameters to the data being analyzed and is not easily scalable. Furthermore, with only nine years of data to work with, one would not expect the statistical procedures to be altogether effective even if they were simple to implement. Ghysels and Ng (2017) shows that unless an ideal adjustment procedure is at our disposable to completely de-seasonalize every series, the imperfectly adjusted data will have a non-negligible common seasonal component upon aggregation. This is particularly problematic if the goal is to extract the common cyclical variations from the data.

While big data means that a large number of series needs to be adjusted, it also comes with a big data blessing: we information in the cross-section dimension. Indeed, if there is commonality in seasonal patterns, it would be inefficient to seasonally adjust each series in isolation. A panel approach is natural though unconventional in the seasonal adjustment literature. Noteworthy is Fok et al. (2007) which controls for seasonal fixed effects using a hierarchical Bayes method to avoid the proliferation of dummy variables needed to control for fixed effects. Instead of treating all dummy predictors as relevant, we train machine learning algorithms to determine which ones to use, and how they are to be used.

3.1 A Two-Step Panel Approach

Our goal is to separate seasonal from non-seasonal sales. In time series analysis, seasonal variations are those that recur with seasonal periodicities. Monthly seasonal variations are those that recur every twelve months. However, as discussed above, weekly variations are

not strictly periodic. We use a definition of seasonality that does not depend on periodicity. Recall that for each state s, we have data at the group and county levels over 469 weeks. Our seasonal adjustment procedure is done for each group g in each state s on a year by year basis using variations across counties and weeks. The analysis is based on the presumption that sales in the same group g collected in different counties share common seasonal patterns over the course of a year.

Some data pre-processing is needed to deal with the fact that some counties are much larger than others, but seasonal patterns over the course of the year should be similar between two neighboring counties even if one county has ten times as many sales as the other. Thus we demean the data year-by-year to remove the size effect. We further standardize the data to ensure scale-independence across years and locations. Normalized sales within each year and each county is defined as:

$$y_{gct} \quad = \quad \frac{\log(\text{SALES}_{gct}) - \mu_{pc\tau}}{\sigma_{gc\tau}}, \qquad \tau = \text{yr}(t)$$

where $\mu_{pc\tau}$ denotes the mean of log sales of group g in county c over the year τ containing week t, and $\sigma_{gc\tau}$ is the corresponding standard deviation. The within-year normalization isolates within-year seasonal patterns while preserving long-term trends in aggregate sales and volatility, which the econometrician can model separately. The normalization also allows us to pool observations across counties in subsequent estimation. Pooling county level data compensates for the relatively short-time span of data for each county.

Next, we posit that y_{gct} has three components: a group specific seasonal component, a common seasonal component, and a cyclical component:

$$\begin{aligned} y_{gct} &= (\text{group specific seasonal sales}) + (\text{common seasonal sales}) + (\text{cyclical sales}) \\ &= d_{gct} + q_{gct} + u_{gct}. \end{aligned}$$

In this decomposition, the seasonal component of sales is $d_{gct} + q_{gct}$. The goal is to extract u_{gct} when only y_{gct} is observed. The point of departure is that (i) d_{gct} and q_{gct} are predictable *over the course of a year*, and (ii) q_{gct} has variations that are common across counties. An overview of the estimation methodology is as follows:

Step 1: Estimation of d_{gct} : For each (g, c) pair, perform least squares estimation of what will be referred to as the Fourier regression

$$y_{gct} = \alpha^0_{gc} + \text{Fourier}(t; \beta_{gc}, \psi_{gc}) + \epsilon_{gct}$$

where with $\delta_{tj} = 2\pi j \frac{\text{day of year}_t}{\text{days in year}}$ and $m_{tj} = 2\pi j \frac{\text{day of month}_t}{\text{days in month}}$

$$\begin{aligned}\text{Fourier}_t &= \sum_{j=1}^{p_d} \beta^s_{k,2j-1} \sin(\delta_{tj}) + \beta^s_{k,2j} \cos(\delta_{tj}) \\ &\quad + \sum_{j=1}^{p_m} \psi^s_{r,2j-1} \sin(m_{tj}) + \psi^s_{k,2j} \cos(m_{tj}) + \text{error}.\end{aligned}$$

The variables δ_{tj} and m_{tj} are strictly periodic albeit non-random. Hence regression R^2 is a summary statistic of the strictly-periodic component of each factor. The regression only includes an intercept and thus preserves any trends in sales that might be in the data. This non-parametric time series regression was used in Pierce et al. (1984) in connection with a stochastic adjustment to seasonally adjust a long time-series of weekly data. The stochastic adjustment can in principle be ARMA modeling, or local linear regressions as in Cleveland et al. (2014). As we only have nine seasonal observations, we exploit information in county-level sales of the same group.

Step 2: Estimation of q^s_{ct} from $\hat{\epsilon}^s_{ct}$: Let $\hat{\epsilon}_{gct} = \widehat{q_{gct} + u_{gct}}$ be the least squares residuals from Step 1. Because Step 1 is based on estimation of a smooth regression, these residuals will have spikes not adequately captured by Step 1. We want to recover the common seasonal component q_{gct} by pooling information across counties. This is done by entertaining a large number of mostly dummy variable predictors. To alleviate the problem of overfitting, we train algorithms to pick out the important ones leaving the functional form of the prediction model unspecified. Details will be explained in the next subsection.

Step 3: Estimation of Seasonally Adjusted Sales Let $\hat{q}_{gct}$ be the prediction of $\hat{\epsilon}_{gct}$ obtained in Step 2. To allow the relative importance of $\hat{q}_{gct}$ and $\hat{d}_{gct}$ to vary across products, we form one final regression

$$y_{gct} = \alpha_{g0} + \alpha_{g1} \cdot \hat{d}_{gct} + \alpha_{g2} \cdot \hat{q}_{gct} + u_{gct} \tag{1}$$

and let the data determine the weights of the two predictable component on y_{gct}. A seasonally-adjusted value of log-sales is obtained by plugging the estimated residual $\hat{u}_{gct}$ into

$$x_{gct} = \widehat{\log \text{adjusted sales}}_{gct} \equiv \hat{u}_{gct} \cdot \sigma_{g\tau} + \mu_{g\tau} \tag{2}$$

This estimate has an intuitive interpretation: the seasonally-adjusted series is the non-predictable part of the series' variation around its overall mean for the year.

The key to our approach is Step 2, a panel-based prediction that involves a large number of predictors and pooling information across counties. We use machine learning methods to predict the common seasonal variations. Subsection 3.2 first discusses the predictors used. The prediction methodology will be discussed in Section 3.3.

3.2 Predictors $\mathbb{Z}_{gct}$

To motivate our modeling choice, consider the event Cinco de Mayo. It occurs on a fixed calendar date, and so is not strictly periodic with respect to the weeks within a year. The seasonal effects of Cinco de Mayo may be more important for counties with a higher Hispanic population. As another example, consider the event of Thanksgiving, which is always on the fourth Thursday in November and is celebrated across the country. The day in the year that Thanksgiving takes place shifts over time. Many of our seasonal events occur at different days of the year. We need a flexible methodology to capture their effects.

With this in mind, we consider date and week specific dummies as well as demographic and spatial predictors collected into $\mathbb{Z}_{gct}$. These are defined as follows. Let START_t denote the date on which week t starts, and END_t denote the date on which week t ends

A. Base case (date-specific) predictors: a dummy variable for each potential calendar date (MM-DD) which is 1 if that date is contained in $[\text{START}_t, \text{END}_t]$ and 0 otherwise. As an example, if $t =$ Feb 4, 2006, the date-specific predictors $\mathbb{Z}^A_{gct}$ is as follows:

01/01	$\cdots$	01/29	$\cdots$	02/04	$\cdots$	12/31
0	$\cdots$	1	$\cdots$	1	$\cdots$	0

B. Week specific predictors.

   i. START_t and END_t's positions within the year (out of 366)

   ii. START_t and END_t's position within their own months (out of 31)

   iii. START_t's position within the month containing END_t (this will be a negative number, and differ from the previous column, if and only if the week ending on t crosses two different months)

   iv. A dummy variable which is 1 if Easter is in the week ending on t, and 0 otherwise

   v. A dummy variable encoding the month in which END_t falls

   For example, for $t =$ Feb 4, 2006, the week -based predictors $\mathbb{Z}^B_{gct}$ will be:

(i)	(i)	(ii)	(ii)	(iii)	(iv)	(v) Jan	(v) Feb	$\cdots$	(v)Dec
28	34	29	4	-2	0	0	1	$\cdots$	0

C. Demographic predictors which depend only on c are drawn from the 2013 American Community Survey, and held constant across time:

i the percentage of the county that is Black, Hispanic, White, and Asian

ii the percentage of the county on SNAP, in poverty, and median household income

iii the percentage of the county c over 60 and under 18

iv Centroid latitude, and centroid longitude for the county

v NOAA's 30-year estimates of average rainfall and temperature for county c during the week of t (which depend on c and t)

The interaction of the three sets of predictors generates as many as 400 potentially relevant predictors. Of these, the 366 date-based predictors form the base case.

3.3 The Prediction Model

Generically denote data with N cases by $\mathcal{D} = (\mathbb{Y}, \mathbb{Z})$ where $\mathbb{Y}$ is the response variable and $\mathbb{Z}$ is the set of observed predictors. To make predictions for all weeks in year τ, we partition $\mathcal{D}$ into $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2)$ where $\mathcal{D}_1$ collects data for all weeks $t \in \text{yr}(\tau)$, and $\mathcal{D}_2$ collects all data not in year τ. The N_1 cases in $\mathcal{D}_1$ will be used for training, and the N_2 cases in $\mathcal{D}_2$ will be used for validation, with $N = N_1 + N_2$. Let z^* be points in $\mathcal{D}_2$.

Since we are interested in predicting the common seasonal variations in the composite error that emerges from the Fourier regression in Step 1, the mapping into $\mathcal{D}$ notation is

$$\begin{aligned} \mathcal{D} &= (\{\hat{\epsilon}_{gct}\}, \{\mathbb{Z}_{gct}\}) = (\hat{\epsilon}^s_{g\tau}, \mathbb{Z}^s_{g\tau}), \quad \forall t \in \text{yr}(t) = \tau \\ \mathcal{D}_1 &= (\{\hat{\epsilon}_{gct}\}, \{\mathbb{Z}_{gct}\}) = (\hat{\epsilon}^s_{g\tau}, \mathbb{Z}^s_{g\tau}), \quad \forall t : \text{yr}(t) \neq \tau. \end{aligned}$$

In words, the training data $\mathcal{D}_1$ consists of observations for all counties in state s over all 469 weeks, less those weeks in year τ (which is 52 except in a leap year). Note that $\hat{\epsilon}^s_{gt}$ is a stacked vector of $\hat{\epsilon}_{gct}$ for all c in state s, and $\mathbb{Z}^s_{g\tau}$ is similarly defined. Thus the training data change with (g, s, τ).

We train algorithms to for prediction group by group, year by year, and state by state. For each state, the exercise involves training $N_{year} \times N_g \times N_{\text{yr}}$ models. For given the predictor set $\mathbb{Z}$, the following models are estimated using training data $\mathcal{D}_1$:

1. POOLED OLS.

2. LARS-type methods.

3. RANDOM FOREST-type methods.

We have close to 400 potential predictors, but we also have $(469 - 52)$ weeks of data for each county. Hence even a pooled least squares regression that uses all predictors (method 1) will have enough degrees of freedom. However, using all predictors is unlikely to be efficient. Hence we consider two machine learning procedures.

Introduced in Efron et al. (2004), the *least angle regression* estimator LARS is functional gradient descent method that repeatedly fits a model to the residuals of the previous step. The LARS solution path is shown to be similar to LASSO as well as boosting and forward stagewise regressions. In fact, these methods have identical solutions under some conditions. Under boosting, each model (also known as learner) is individually weak, but is 'boosted' to produce a strong learner via averaging. Averaging in this case reduces bias. The learner can be but not restricted to a regression tree. The boosting perspective is helpful to understand the difference with the method of random forests. Our implementation of LARS-type methods is actually based on LASSO because it requires fewer choices of tuning parameters.

Random forest (RF) due to Breiman (2001) is an ensemble method that builds a prediction from a collection of regression trees. Each tree is fitted to a randomly selected subset of predictors in a bootstrapped sample. The final model is also an average over the trees, but the trees are built either separately or in parallel rather than sequentially as is the case with LARS type estimators. Regression trees can uncover complex relations and are strong learners, but they tend to have have high variance. Averaging in this case reduces the variance of models that have low bias. Random forest is an extension of BAGGING which averages over trees grown on bootstrapped samples using all predictors.

Both LARS and random forest-type predictions are based on averaging over small models that use only a subset of available predictors. Hence they are more resistant to overfitting. Though these methods have been widely applied to iid data, applications to time series data are more limited. Success of these algorithms in the present setting is very much an empirical matter. Random forest is the most flexible since it does not impose linearity or smoothness. We use it as benchmark in the discussion of results.[6]

[6]We implement random forests using the R package RANGER with default parameter settings. That is , the default size of forest is —ntree—=500 trees, and the default value of —mtry— (the number of independent variables considered for each split) is the square root of the total number of independent variables. The —min node size— parameter, which controls the depth of each tree grown, is set to 5 by default.

4 Analysis of the Seasonally Adjusted Data

We begin with a broad overview of the adjusted data. The share of a group in the seasonally adjusted budget is now the ratio of seasonally adjusted sales for the group to total adjusted sales summed across groups. Comparing the shares from the raw data and the adjusted data, we note that the largest differences are in groups like 'floral', 'insecticides', canning, ice, fragrances, toys, stationary, and candies'. These results make sense since outdoor activities are concentrated in the summer, stationary spending is highest around 'back to school' time, while spending on candies and fragrances tend to be higher around Valentine's Day.

The premise of our analysis is that residuals from the univariate Fourier regressions have comovements that are predictable. To evaluate the incremental predictive power provided by different adjustment methods, we consider the R^2 corresponding to (1), which is a regression of log sales y_{gct} on the two estimated seasonal components: $\hat{d}_{gct}$ and $\hat{q}_{gct}$. Table 1 summarizes the distribution of R^2 over all groups and states. Roughly 50% of the variations in log-sales are seasonal and predictable. The degree of predictability varies across groups, ranging from 14% to over 90%. Notably, Step 2 improves upon the univariate Fourier regressions implemented in Step 1 alone. Since the highest and lowest quantiles of the R^2 do not depend on the procedure, the improvements apply not to a few groups with extreme seasonality, but to a large number of groups.

Table 1: Importance of Seasonal Component

	Average of R^2 in Equation (1)					
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
RF	0.54	0.55	0.91	0.68	0.39	0.14

The random forest procedure is the most flexible and not surprisingly captures the highest predictability, though LASSO also achieves improvements. Which machine learning method to use is of second order importance. More important is to exploit cross-section information via Step 2. Figure 2 illustrates the difference by plotting the R^2 of random forest against that based om the Fourier method. If the random forest estimator provides relatively little additional information, the third-stage regression will push λ_g towards zero. In such cases, the R^2 values will be bunched along the 45 degree line. Figure 2 indicates such groups do

exist. However, many other groups have values in the scatterplot located above the 45 degree line. For some of these groups, the improvement in fit from adding the panel data step is quite significant. A quarter of groups see R^2 increases of 13% or greater.

4.1 Demand Analysis of Seasonally Adjusted Data

At face value, it may seem that the improvement of a few percentage points in predictability over the univariate Fourier regression is trivial. However, the data adjusted by machine methods have far fewer spikes than those adjusted using the Fourier regressions alone. This difference has direct implications for demand estimation. We find four factors in the data adjusted by the Fourier Step alone. The first two factors explain over 68% of the variation in the data and consists of a trend and a cyclical component. However, factors three and four remain spiky and quasi-periodic, indicating that the Fourier regressions by themselves leave non-trivial seasonal variations unexplained.

The criterion of Bai and Ng (2017) finds either three or four factors in the shares adjusted using the machine learning methods depending on the state. Compared to factors estimated from the raw data, the most notable difference is absence of large spikes. This is consistent with the finding that the spectrum of the adjusted factors no longer display spikes around .1208j, which corresponds to a periodicity of around 52 weeks.

Figure 3 plots the three factors in ALL based on random forests. These will be denoted $\hat{F}_{\text{RF}}$ to distinguish them from the ones estimated from the unadjusted data, now denoted $\hat{F}_{\text{NSA}}$. The first three adjusted factors together explain about 80% of the variations of the adjusted data, with $\hat{F}_{1,RF}$ explaining 56% and $\hat{F}_{2,RF}$ explaining 15%. As can be seen from Figure 3, $\hat{F}_{1,RF}$ is highly persistent and has a trend component. An investigation into the factor loadings finds that $\hat{F}_{1,RF}$ always loads heavily on books and magazines, ethnic hair treatment, and photographic supplies. These are groups that have likely experienced secular changes in demand during our sample period.

Even though the Nielsen data is concentrated on grocery store sales with few consumer durables that are traditionally known to be cyclical, $\hat{F}_{2,RF}$ is clearly cyclical and warrants further investigation. The series is also persistent and its spectrum has mass piled up around frequency 0.05, which corresponds to periodicity of 125 weeks, or 2.5 years. We use two consumer confidence index as benchmark of cyclicality:- the Rasmussen RCCI index and the Bloomberg index of consumer confidence. The former is a daily national survey collected by the Rasmussen grop that tracks 1,500 consumers concerning their confidence, expectations, and sentiment toward the U.S. economy. The latter started as the ABC News consumer

comfort index and was under control of the Bloomberg corporation since 2011. Figure 4 plots $\hat{F}_{2,\text{RF}}$ (in dark green) along with the Rasmussen index RCCI (in blue) and the Bloomberg consumer confidence index (in red). These confidence measures have absolute correlation with $\hat{F}_{2,\text{RF}}$ of about 0.8. Though not evident, $\hat{F}_{2,\text{RF}}$ is is strongly correlated with $\hat{F}_{4,\text{NSA}}$. A regression of $\hat{F}_{4,\text{NSA}}$ on $\hat{F}_{2,RF}$ yields an R^2 of 0.6. The largest residuals of that regression are precisely spikes between weeks 46 and 50, indicating that Step 2 is picking up the spikes that the Fourier regressions could not account for in Step 1.

Turning now to the third adjusted factor, $\hat{F}_{3,RF}$ takes a big dip in week 12, ie the week ending March 22, of 2008. As point of reference, JP Morgan purchased Bear Stearns on March 17 of 2018. Furthermore, oil price spiked up to nearly 110 dollars per barrel during that week. The factor is actually highly correlated with the 52 week change in consumer confidence. Figure 5 plots $\hat{F}_{3,RF}$ in ALL with the 52 week change in RCCI and Bloomberg. The correlation of $\hat{F}_{3,RF}$ are 0.74 and 0.68, respectively. If $\hat{F}_{2,RF}$ is a coincident index, $\hat{F}_{3,RF}$ is informative about direction of change in consumer sentiment, and we refer to it as a 'direction factor'.

How does this weekly information relate to the official monthly retail sales? The U.S. Census Bureau releases both the raw and the seasonally adjusted each month.[7] We fix the values during weeks in the same month at the officially released values. Figure 6 plots both series along with $\hat{F}_{2,RF}$. The top panel shows that our $\hat{F}_{2,RF}$ has a correlation of 0.65 with the officially adjusted series. The bottom panel plots the sames in terms of 52 week changes. The correlation of the adjusted series is 0.74. The most notable difference is around the 2008 Financial Crisis. The $\hat{F}_{2,RF}$ shows a steeper decline than the official data, but the weekly series tracks the monthly series reasonably well.

4.2 Sensitivity to Aggregate Conditions

A unique feature of the scanner data is the information at the spatial and group levels at the weekly frequency. This presents an opportunity to study response of spending to economic conditions at a disaggregated level. We begin with an analysis at the group level.

The marginal effect of the $\hat{F}_{2,\text{RF}}$ on the share of group j is given by ithe loading on the factor, but the total effect depends also the share j. To get a unit free measure of the importance $\hat{F}_{2,\text{RF}}$, we compute $\text{SGN}R_j^2$, which is the fraction of variance of share j explained by $\hat{F}_{2,\text{RF}}$, where SGN is the sign of the loading of $\hat{F}_{2,RF}$ on share j. Since $\hat{F}_{2,RF}$ is positively correlated with RCCI, a positive value indicates that share j is pro-cyclical, while a negative

[7]The series are RETAILSMNSA and RETAILSMSA in FRED.

value means that share j is high when $\hat{F}_{2,\mathrm{RF}}$ is low. Figure 8 plots the distribution of $\mathrm{SGN}R_j^2$ for ALL to illustrate the dispersion of sensitivity to aggregate conditions across product-groups. It is asymmetric with more counter-cyclical groups than pro-cyclical ones. For NY and ALL, the statistical effect of an increase in $\hat{F}_{2,\mathrm{RF}}$ is strongest on liquor. For FL and TX, the largest positive impact is on cosmetics. For CA, the largest positive impact is on prepared foods. These are arguably luxury goods. There are also goods that are little affected by $\hat{F}_{2,RF}$. Included in this list are disposable diapers, shaving products, cold and cough remedies, and somewhat surprisingly, beer.

As for the effect of a decrease in $\hat{F}_{2,RF}$ in CA, FL, TX, and ALL generates the strongest increase in frozen and canned vegetables. In NY, the effect is strongest on pasta. This suggests less eating out during downturns in favor of preparing meals at home. To further explore this possibility, we take the five groups with the highest loadings on $\hat{F}_{2,\mathrm{RF}}$: frozen vegetables, canned vegetables, pasta, frozen pizza/snacks, and condiments/sauces. Because these all seem to be food groups related to home-cooking, we designate them the"food-in" group. For each state, we compute the share of food-in. The series is most persistent in CA and least persistent in NY, with first order autocorrelation coefficientn of around 0.75 and 0.50 respectively.

Figure 9 shows the response of food-in to a one standard deviation shock in $\hat{F}_{2,RF}$ in a VAR consisting of food-in, $\hat{F}_{2,RF}$, $\hat{F}_{3,RF}$ and RCCI. A positive shock, which is an increase in economic activity, has a statistically well determined effect on food-in that is negative and peaks after two weeks. Though food-in rebounds after five weeks, there are lingering negative effects. About 44% of the variations in food-in are explained by its own lag, 48% explained by $\hat{F}_{2,RF}$, 8% by $\hat{F}_{3,RF}$ with little attributed to RCCI. The results are robust to whether RCCI is ordered second or last. Interestingly, even though the correlation between RCCI and food-in is well over 0.75, shocks to RCCI account for little of the variations in food-in. This suggests that while RCCI is a valuable economic indicator, it has limited structural information.

The impact of an increase in $\hat{F}_{3,RF}$ tends to be most negative on eggs and most positive on dried-fruit, which is often marketed as a snack.

To be completed.

Appendix: Random Forest

Random forest can be thought of as a histogram estimator in which the support of the predictors is partitioned into bins, and the conditional expectation in the bin will be approximated by the mean of the mean. A sketch of the algorithm is as follows. There are four main parameters: the number of bootstrap samples, n, the number of trees, `ntrees`, the number of parameters to sample, `mtry`, and the minimum node size, `min_node_size`. To generate a prediction $\hat{f}^m$ for $m = 1, \ldots$`ntrees`:

a. Sample n cases $\mathbb{D}^m = (\mathbb{Y}^m, \mathbb{Z}^m)$ from the training data $\mathcal{D}_1$ with replacement.

b. Grow a regression tree to predict y^m given z^m. At each "branch" of the regression tree, consider splitting on `mtry` randomly-sampled components of z^m, picking the component whose split best explains the data

c. Stop growing the regression tree once there is a "leaf" containing less than or equal to `min_node_size` training observations

Once `ntrees` predictors have been grown, we let the ensemble predictor be

$$\hat{f}(z) \equiv \frac{1}{\texttt{ntrees}} \sum_{m=1}^{\texttt{ntrees}} \hat{f}^m(z)$$

In our setup, we seek a prediction $f = q^s_{g\tau}$ for each product g (e.g., beer), state s (e.g., New York), and year τ (e.g., 2009). We use random forest to train a model for each (group, state, year) triplet by pooling cross-sectional information across counties, and time series information across weeks other than the year τ in which a prediction is to be made.

While the random forest is regarded as one of the best prediction algorithms available, establishing the statistical properties of random forests has been difficult because of the complex interaction between the sampling scheme, tree splitting, and tuning parameters. There are few guides to choose tuning parameters but they appear to have limited effects on output in empirical evaluations. Breiman (2001) uses as loss function the difference in variance before and after an additional split and sets $p = [\dim(\mathbb{Z})/3]$ what is p? Use $n_{??}$ skimming the Breiman paper I didn't find any mention of this–should we just remove this sentence? for regressions. The consistency proofs available are based on algorithms that are somewhat different from the implementations.[8]

[8]For the statistical theory of random forests, see discussions Buhlmann and Yu (2002), Scornet et al. (2015), Biau and Scornet (2016) and Scornet (2016).

References

Bai, J. and Ng, S. (2017). Regularized estimation of approximate factor models. mimeo, Columbia University.

Banks, J., Blundell, R. and Lewbel, A. (1997). Quadratic engel curves and consumer demand, *Review of Economics and Statistics* **79**: 527–539.

Biau, G. and Scornet, E. (2016). A random forest guided tour, *Test* **25**: 197–227.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Buhlmann, P. and Yu, B. (2002). Analyzing bagging, *Annals of Statistics* **30**(4): 927–961.

Cleveland, W., Evans, T. and Scott, S. (2014). Weekly seasonal adjustment- a locally weighted regression approach. U.S. Bureau of Labor Statistics Working Paper 473.

Deaton, A. S. and Muellbauer, J. (1980). An almost ideal demand system, *American Economic Review* **70**: 312–326.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics* **32:2**: 407–499.

Fok, D., Franses, P. and Paap, R. (2007). Seasonality and non-linear price effects in scanner-data-based market response models, *Journal of Econometrics* **138**: 231–251.

Ghysels, E. and Ng, S. (2017). Seasonality and big data. mimeo.

Gorman, W. M. (1981). Some engel curves, *in* A. Deaton (ed.), *Essays in the Theory and Measurement of Consumer Behavior in Honor of Sir Richard Stone*, Cambridge University Press.

Harvey, A., Koopman, S. and Riana, M. (1997). The modeling and seasonal adjustment of weekly observations, *Journal of Business and Economic Statistics* **15**: 354–368.

Lewbel, A. (1991). The rank of demand systems: Theory and nonparametric estimation, *Econometrica* **59:1**: 711–730.

Lewbel, A. (1997). Consumer demand systems and household equivalence scales, *in* M. H. Pesaran and P. Schmidt (eds), *Handbook of Applied Econometrics*, Vol. 2:4, Blackwell, Oxford, pp. 167–201.

Lewbel, A. (2003). A rational rank four demand system, *Journal of Applied Econometrics* **18**: 127–135.

Muellbauer, J. (1975). Aggregation, income distribution, and consumer demand, *Review of Economic Studies* **62**: 269–283.

Ng, S. (2017). Opportunities and challenges: Lessons from analyzing terabytes of scanner data, *in* B. Honore, A. Pkes, M. Piazzesi and L. Samuelson (eds), *Advances in Economics and Econometrics, Eleventh World Congress of the Econometric Society*, Vol. II, Cambridge University Press, pp. 1–34.

Pierce, D., Grupe, M. and Cleveland, W. (1984). Seasonal adjustment of the weekly monetary aggregate: A model based approach, *Journal of Business and Economic Statistics* **2**: 260–270.

Scornet, E. (2016). On the asymptotics of random forests, *Journal of Multivariate Analysis* **146**: 72–83.

Scornet, E., Biau, G. and Vert, J. (2015). Consistency of random forests, *Annals of Statistics* **43**(4): 1716–1741.

Figure 1: Factors Estimated from Raw Shares: All States

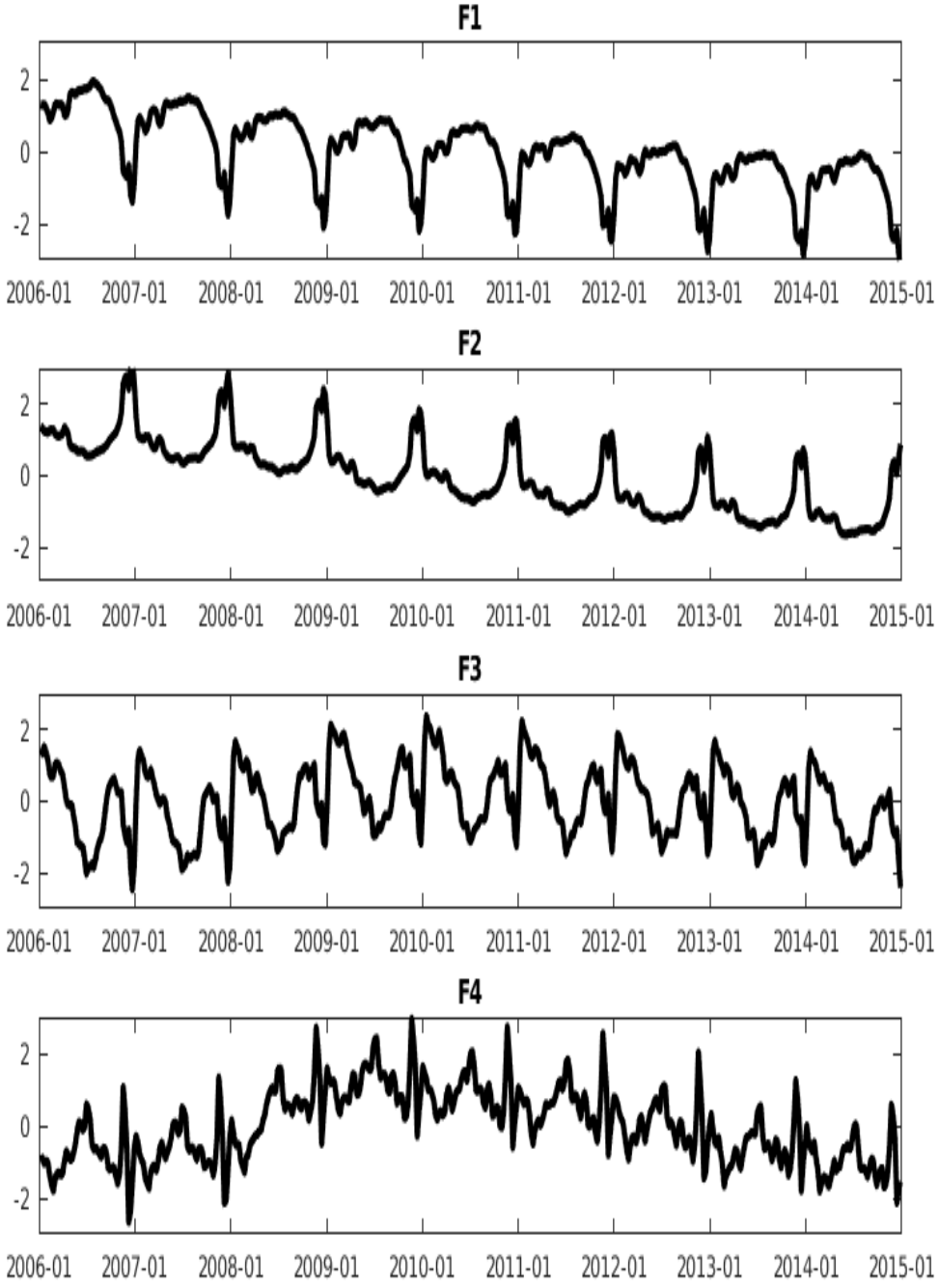


Figure 2: Incremental Predictive Power of Random Forests

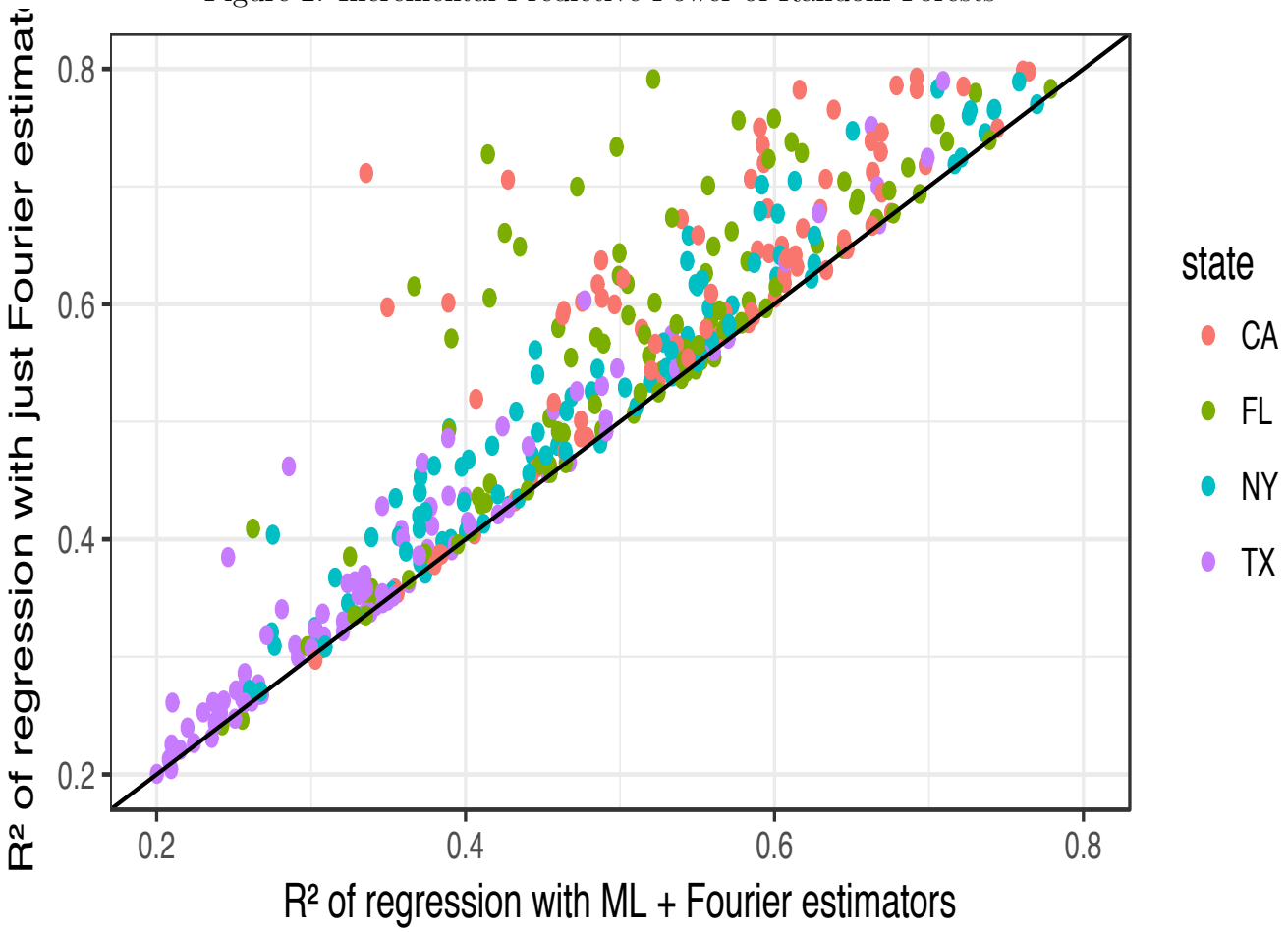


Figure 3: Factors Estimated from Seasonally Adjsted Shares: All States

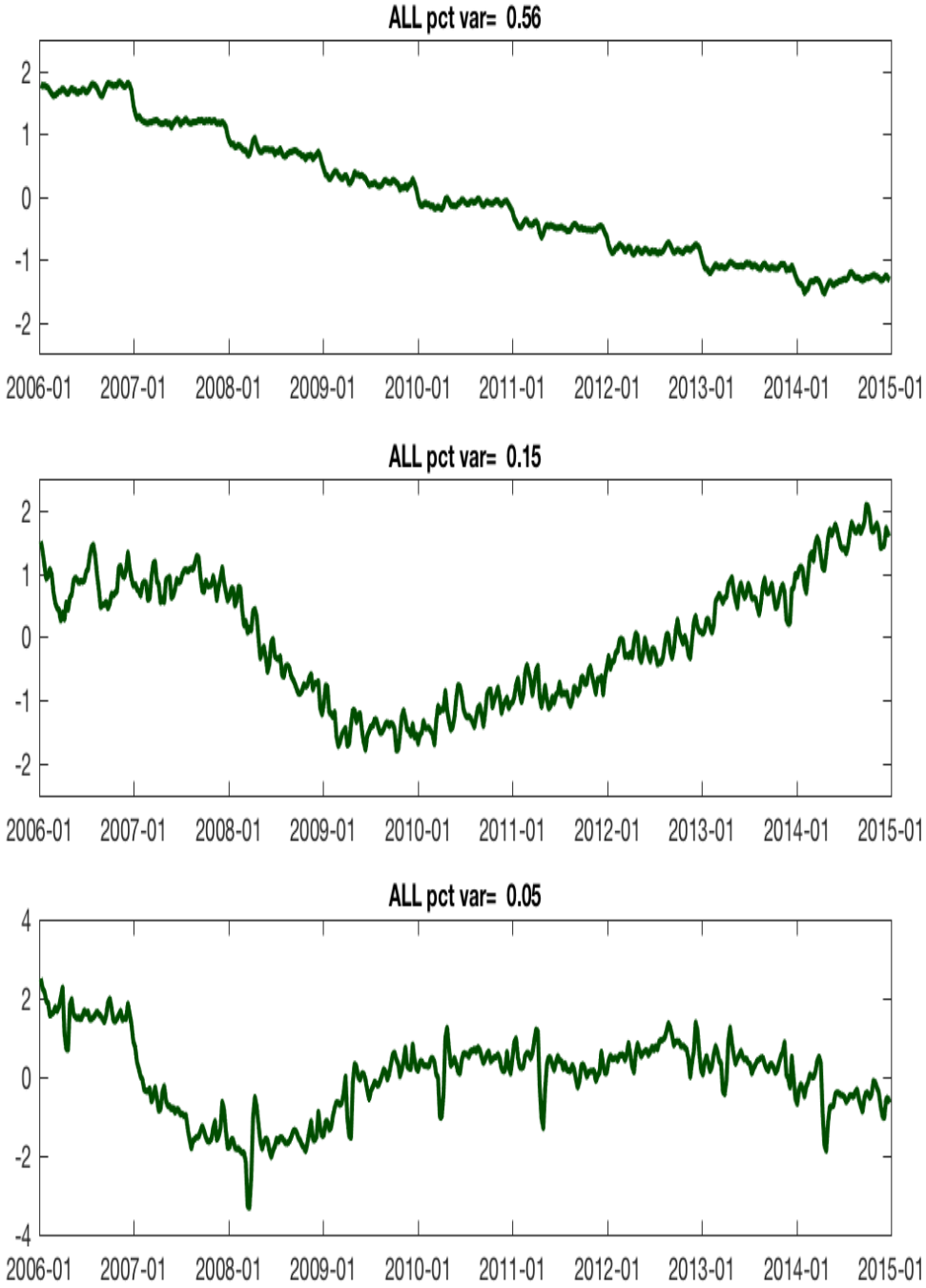


Figure 4: The Coincident Factor: $\hat{F}_{2,RF}$

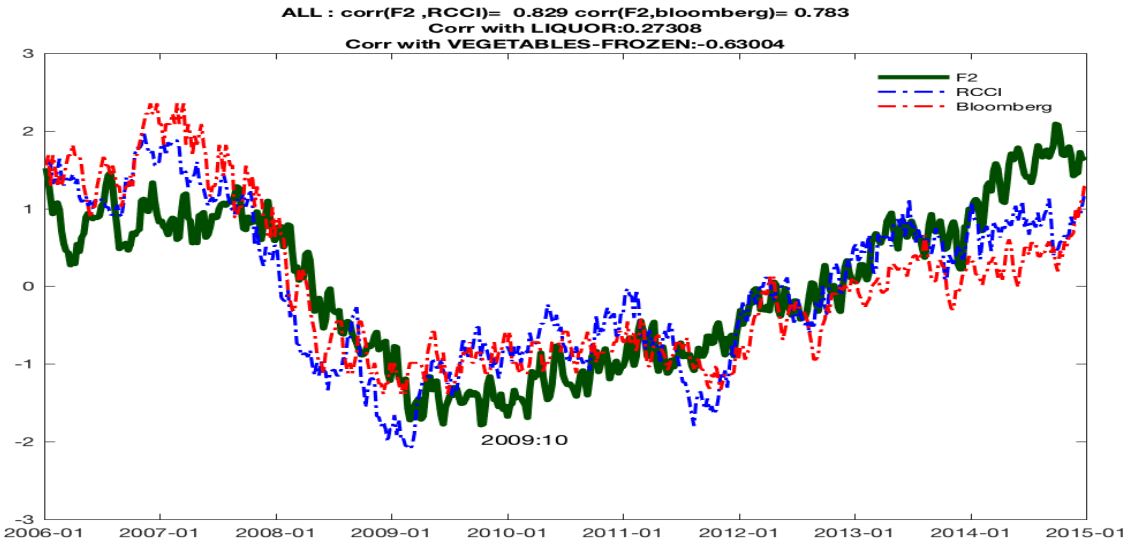


Figure 5: The Direction Factor: $\hat{F}_{3,RF}$

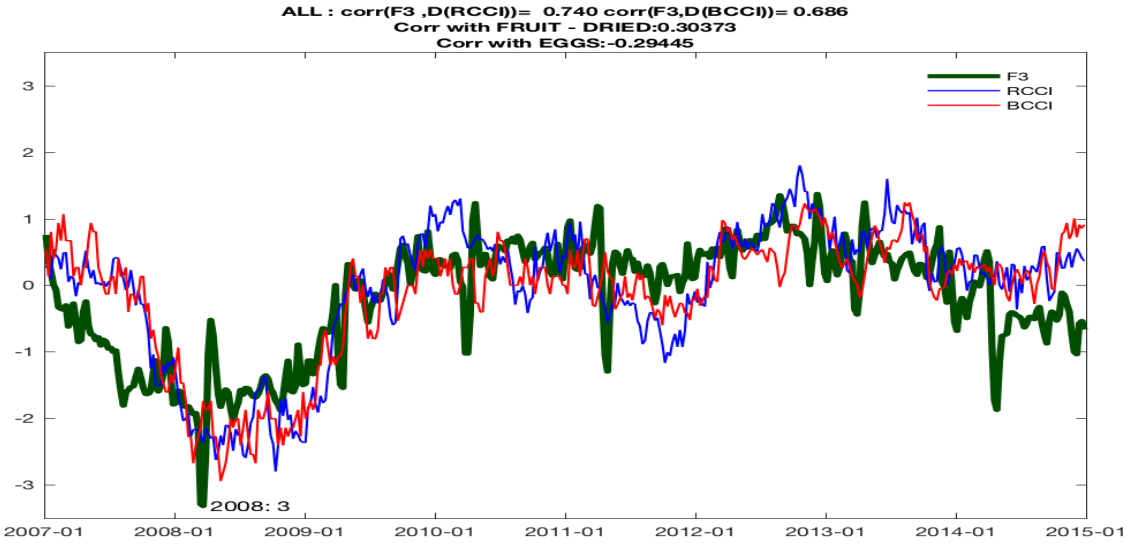


Figure 6: Comparison with Monthly Retail Series

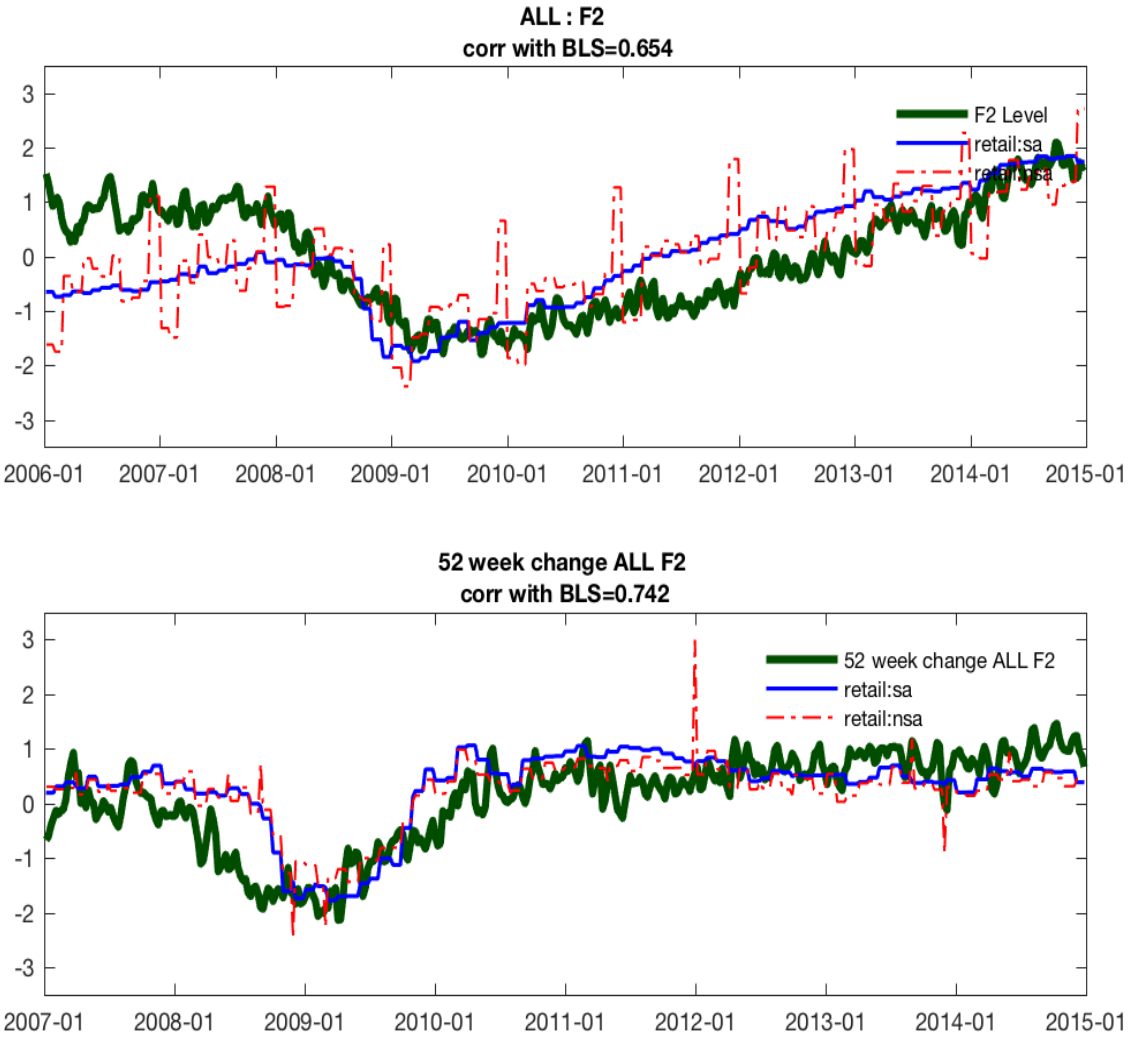


Figure 7: R^2 of adjusted shares on $\hat{F}_2$: ALL States

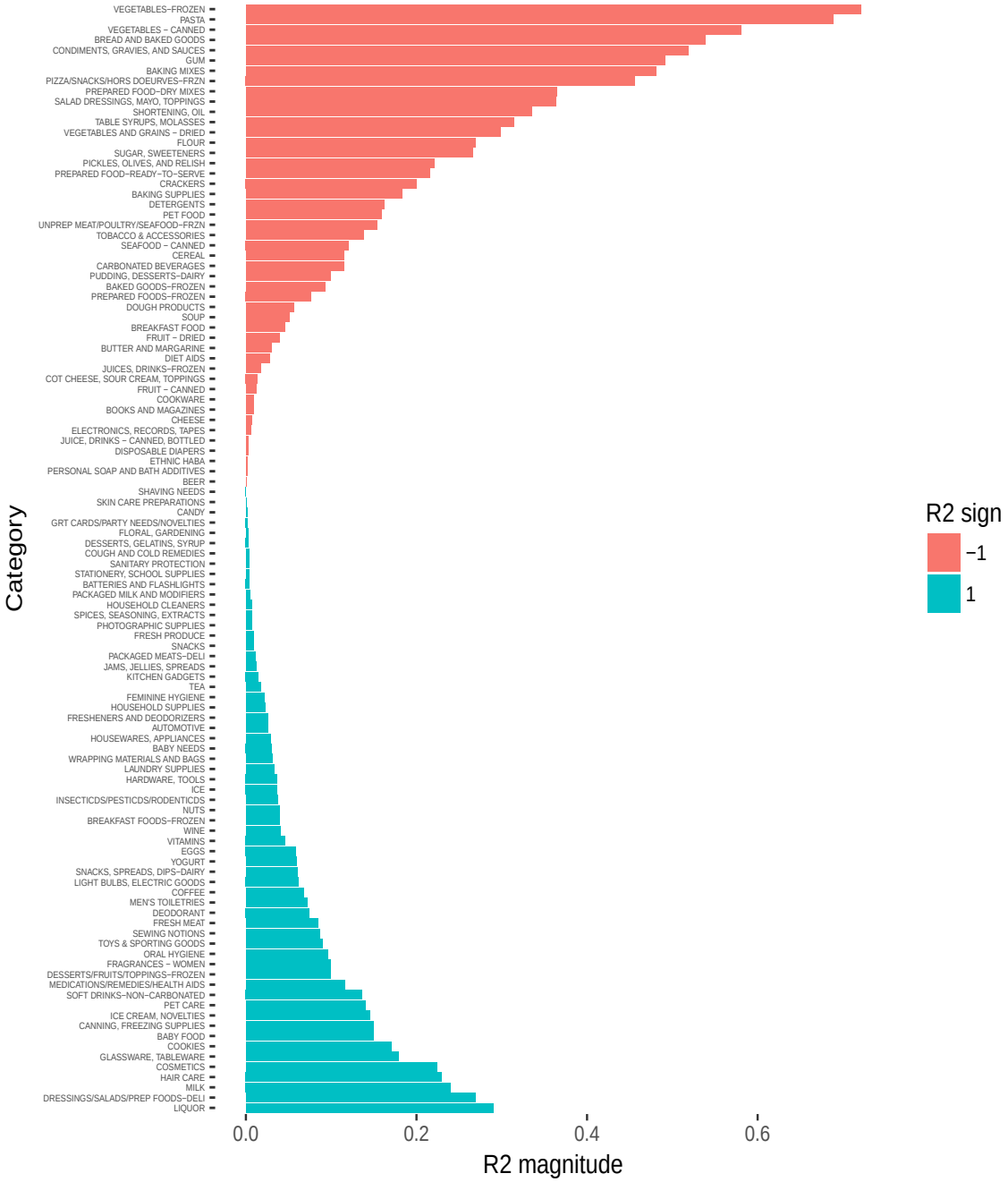


Figure 8: $\hat{F}_2$ loadings on product shares: ALL States



Figure 9: Response of food-in

