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

Rishab Guha*
Harvard University

Serena Ng†
Columbia University and NBER

March 18, 2018

Abstract

This paper analyzes weekly sales of 108 categories of consumer products 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 remove seasonality 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 products. We treat seasonality as a prediction problem and use machine learning methods to remove product and spatial fixed effects. A demand system estimated using seasonally adjusted sales has a rank of three, implying that Engel curves are spanned by three functions of prices and income that are common across goods. The factors seem to have business-cycle interpretations, and factor loadings imply a cross-product distribution of recession sensitivity. The seasonally-adjusted share vector also seems to hold predictive information about county-level unemployment. The scanner data can thus be a rare source of timely information about local economic conditions, provided that one can filter out seasonal, and other deterministically-predictable, effects.

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.

Extended Abstract

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 are characterized by what data scientists refer to as the "three Vs": volume, variety, and velocity. Econometricians may think of them as short panels of big, often unbalanced, high-frequency, highly heterogeneous data. Such data can potentially allow for more precise and detailed analyses of micro-level behavior and aggregate economic phenomena than tidier traditional datasets. However, as discussed in Ng (2017), fully exploiting all the data—as opposed to using subsets or aggregates—also brings unique challenges. A case in point is the Nielsen Retail Scanner data, which records sales of consumer goods at a week/store/UPC level of granularity.

The Nielsen data have three features of interest for economic analysis. First, the data have records of real-time sales and unit prices collected at the store level. Such transactions data are distinctly different from price indices published by government agencies. Once transactions are recorded, they are not subject to revisions. Second, the data are available across spatially diverse regions on a weekly basis. This offers a snapshot of spending behavior that is different from the official retail sales data that are aggregated across regions and over weeks. Third, the 2006-2014 sample includes a big recession, a memorable hurricane, and a government shutdown. The Nielsen data are heavily weighted towards groceries and mass-merchandise products with limited coverage of durable goods. However, in principle, much can still be learned from our panel of weekly sales data.

Moving from principle to practice is challenging because the data is highly heterogeneous in the product and spatial dimensions, while also exhibiting pervasive cyclical and seasonal fluctuations. It is perhaps best described as a short T panel of data with multi-dimensional fixed effects, and cyclical variations that are convolved with strong seasonal effects. It is quite common in business cycle research to aggregate this data from weekly to a monthly or quarterly frequency to mitigate the problem of seasonality, but this comes at the expense of surrendering the weekly information that makes the data unique.

This paper takes two approaches to extracting economically-interesting information from the disaggregated data. The first isolates the cyclical and seasonal variations that are common across products. This is accomplished by analyzing time-variation in budget shares from the perspective of a demand system. In demand analysis, the common variations across products are spanned by latent functions of prices and income. Our demand system

consists of observations for 108 products over 469 weeks. This allows us use methods in large dimensional factor analysis to estimate the common factors from budget shares, which are the functions of prices and income that span Engel curves, without directly using data on prices and income. The estimated common factors shed light on the relative importance of the common seasonal and cyclical component in budget shares.

Using scanner data for California (CA), Florida (FL), New York (NY), and Texas (TX), and the criterion of Bai and Ng (2017), we estimate demand systems with a rank of four to five. Previous empirical studies using seasonally-adjusted data, such as Lewbel (2003, 1997); Banks et al. (1997) have estimated demand-system ranks of two to four, and Gorman (1981) shows that exactly aggregable demand systems must have rank no larger than three. Investigation of the identified factors suggests that the two dominant factors are exclusively seasonal, highlighting the importance of seasonal adjustment for economic analysis.

Our second approach tackles the challenge of seasonally adjusting short, complex, weekly, data head-on. Our point of departure is the observation that while spatial, product, and time of the year effects contribute to heterogeneous seasonal variations, much of this variation is highly predictable *ex-ante*. Therefore, we treat seasonal adjustment as a prediction problem. Our main contribution is a two step procedure. First, we use traditional regression analysis to remove purely periodic variations at the county/product level. In the second step, the county-level data for each product are pooled at the state level. We then use a machine-learning approach (primarily random forests) to isolate seasonal variation that is cross-sectional, or not strictly periodic. The method differs fundamentally from existing univariate regressions or filtering methods which rely long spans of the data. By letting algorithms perform estimation and variable selection, we by-pass the need to specify a single data generating process, which is a difficult task when the data are highly heterogeneous, and seasonal effects are not strictly-periodic (as is often the case with weekly data).

Our analysis yields five interesting observations. First, most of the variation in the data is seasonally predictable. On average, about 55% percent of the overall variation in product sales is predictable. The result is robust across geographical regions. Second, for many products a substantial amount of the predictable seasonal variation is not strictly-periodic. One class of non-periodic predictable variation comes from holidays whose "weekly" position within the year moves around over time. Another can be traced to interactions between time of the year and variables related to store location. Our random-forest predictor is shown to capture information omitted by linear regression models and classical seasonal adjustment methodologies. Third, even within the class of non-durable consumer goods contained in the

retail scanner data, there is substantial product-level heterogeneity in response to economic stress. While the sales of some products are little affected by local conditions, other products actually benefit from local economic distress. We document a 'distribution' of recession sensitivity that may be of interest in monitoring exercises. Fourth, we find that during the Great Recession, the seasonally-adjusted budget-share vector at the county level has predictive power for the one-month ahead change in that county's unemployment rate. This is interesting because few economic indicators are available at the local level, and to our knowledge none are available at a weekly frequency. Our analysis suggests that once properly adjusted the weekly scanner data can play this role. Fifth, we re-examine the interaction between the seasonal and the business cycle, and find that they are mostly uncorrelated: the level of seasonal predictability does not seem to co-move with unemployment or uncertainty cross-sectionally or in the time-series.

1 To do

An incomplete draft of 20+ pages with preliminary results is available at `www.columbia.edu/~sn2294/papers/guhang.pdf`

We plan to further develop our research by extending our seasonal-adjustment methodology to the task of seasonally-adjusting high-frequency price observations. This will let us test a more granular version of Barsky and Miron (1989)'s observation that in the seasonal cycle, as in the business cycle, prices vary less than quantities. With seasonally-adjusted prices, we will also be able to integrate our data into a large-scale high frequency demand system along the lines of the Almost Ideal Demand System proposed by Deaton and Muellbauer (1980). With seasonally-adjusted granular data, we will also be able to examine the immediate reaction of consumer spending to extreme weather events or policy shocks (e.g., the 2013 federal government shutdown); with unadjusted data, any response is swamped in seasonal noise. should we include a sentence saying that we also plan to expand the analysis to states other than CA, FL, NY, TX. This allows us to evaluate differences in demand patterns across states.

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.

Barsky, R. and Miron, J. (1989). The seasonal cycle and the business cycle, *Journal of Political Economy* **97**(3): 503–534.

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

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.

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.

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.