Extracting Latent Economic Signal from Online Activity Streams

Joseph Reisinger
The University of Texas at Austin
joeraii@cs.utexas.edu

Abstract

Online activity generates a myriad of weak economically relevant signals, ranging from content-rich interactions such as blog posts, venue reviews and status updates, to more transient or even autonomic signal such as search queries, purchases or realtime ad auctions. In this work we develop an approach to distilling latent economic indicators from large collections of realtime activity streams. In particular we investigate the relative contribution of several online activity streams on forecasting a large cross-section of economic indicators. In order to effectively overcome noise inherent in the extracted weak signals and prevent overfitting, we develop a low-rank vector autoregression (VAR) formulation based on sparse matrix factorization. Large-scale mining of online activity data can potentially augment traditional economic analyses in several ways: (1) *improving forecasting*: E.g., we find social network factors are more predictive of unemployment and consumer discretionary spending than a basket of econometric factors, (2) *capturing network effects* via cross-domain *social transfer*: E.g., increases in unemployment indicators lead to technology and computing-related searches; wedding planning on Yelp is a leading indicator of positive sentiment on Facebook, and (3) the development of macroeconomic and sentiment indicators with significantly *lower latency* than is traditionally available.

1 Background

Over the last decade, the Web has emerged as the primary medium for a large number of economic transactions, expression of consumer, producer and political sentiment, and the development novel trends [cf. 15]. Goel et al. summarize this observation succinctly, regarding online activity as "a snapshot of the collective consciousness, reflecting the instantaneous interests, concerns, and intentions of the global population" [14]. In this work, we will use the term *online activity* to refer to the broad spectrum of events resulting from users interacting with the Web. Such events range from the content-rich, e.g. posting news articles, blog entries, product and venue reviews, and status messages on social media sites, to the more transient, e.g. Web searches, location-based venue checkins, page views and ad impressions.

As proxies for consumer behavior, online activity streams have been posited to predict a wealth of economic, social and political indicators. For example, Gruhl et al. use blog "chatter" to predict the sales rank of items on Amazon [15]; O'Connor et al. use Twitter sentiment to predict public opinion around elections [23]; several studies use Twitter to predict the box-office success of movies, e.g. via sentiment clues [20] or simply tweet volume [2]; and Google uses flu-related search queries to predict outbreaks, often faster than the CDC [12].

Online sentiment has also been shown to be informative for econometric problems. For example, Tetlock finds that negative media sentiment can forecast short term movements in stock price [25]. Gilbert and Karahalios find that collective mood is potentially predictive of financial decisions [11].

Google maintains a *domestic trends* index [1] tracking economically relevant aggregate search traffic across a diverse set of sectors, ranging from auto and home sales to US unemployment claims [6]. Finally the *Billion Prices Project* seeks to define a bottom-up measure of CPI based on extracting and classifying prices from online retailers [5].

We contribute to this notion of "Web as Economic signal" in two ways: (1) extracting indicators of online activity from a large variety of sources, and (2) developing a robust modeling approach based on vector autoregression (VAR), bridging the gap between large-scale Web extraction and traditional econometric analysis. VAR models are used to capture symmetric linear interdependencies between multiple time series without imposing significant model constraints.

In order to address overfitting and spurious correlations due to high dimensional data, we extend VAR by assuming a a low-rank ℓ_1-sparse autoregressive factor structure. This *sparse pooling VAR* (spVAR) model can be implemented efficiently using Forward Backward splitting [7], allowing it to robustly scale to upwards of 10k base time-series. spVAR is similar to dynamic factor models for large cross-sectional VARs [cf. 9, 24], where it is assumed that dynamic interrelations between variables can be explained by a few common factors. However, in this work we assume factor stationarity since our online activity sample consists of only a single year.

We apply spVAR to a comprehensive set of indicators derived from both econometric data and online activity including search queries, venue reviews, microblogs, status updates and realtime ad prices. A total of 133 separate signals (116 based on Web activity and 17 econometric) are included in this study. Due to the large number of hypotheses tested, in order to control for inflated false positive rate due to multiplicity, we adjust p-values using *false discovery rate* [8].

We find that social network factors are more predictive of unemployment and consumer discretionary spending than a basket of econometric factors. Furthermore this effect is robust across disparate data sources, e.g. holding true for both Yelp reviews and Twitter posts, lending evidence for the existence of a low-rank set of social factors influencing economic variables. Hence, incorporating measures of online activity can potentially augment traditional economic analysis.

2 Data Sources

We collect 12 months of daily Web activity data (2010-10-18 to 2011-10-18) from six high-dimensional, low-latency sources:

- **Twitter** – A popular microblogging service containing news content and public mood.
- **Yelp Reviews** – Review data from the crowd-sourced local business site Yelp.
- **Crowdsourced News** – Article headlines from the top stories posted to the tech blog Hacker News.
- **Facebook Gross National Happiness** – Aggregate sentiment indices derived from Facebook status message updates [**fb_gnh**; 21].
- **Ad Exchanges** – Supply-side display ad pricing data for 11k publishers (10B impressions) on three large ad exchanges **ad_ex**.
- **Google Domestic Trends** – Daily relative search term frequency across a wide range of economic and topical sectors (e.g. Advertising, Construction, Durable goods, Mortgage, and Unemployment) [**gdi_us**; 6].

Raw data from Twitter, Yelp and Hacker News are converted into low-dimensional time series by using Latent Dirichlet Allocation [4] with K=100.

In order to place the online activity stream data in a broader context, we compare it with a set of macroeconomic time series from the Federal Reserve Data research archive (collectively, the **econometric** set)[1][2] and a set of financial time series collected from Yahoo finance[3]

[1] research.stlouisfed.org

[2] (**Data Granularity**) We focus on high-granularity series with either daily or weekly availability as several of our online activity series are only available back for a single year. This precludes richer economic data such as CPI, GDP, consumer sentiment and housing starts which are available monthly or even quarterly.

[3] finance.yahoo.com

The primary macroeconomic indicators we collect are *M2 Money Supply* and *Weekly Initial Jobless Claims*. The M2 money supply measure (**fred_m2**) is commonly used in macroeconomic models to forecast inflation [16]. Jobless claims tracks "emergent unemployment" via the number of people filing for unemployment insurance each week. In particular we are interested in whether changes in money supply or unemployment result in shifts in online behavior (or vice-versa). We also collect data on, e.g., large cap stocks (**fred_sp500**), near term volatility or "fear" (**fred_vixcls**) and corporate bond yields (**fred_dbaa**).

From Yahoo, we include market sectoral indicator proxies available as publicly traded ETFs: Energy (**XLE**), Financials (**XLF**), Industrial (**XLI**), Technology (**XLK**), Utilities (**XLU**), Consumer Discretionary (**XLY**).

3 Sparse Pooling Vector Autoregression

We propose an approach to structural modeling and forecasting based on the *vector autoregressive* (VAR) process [17] that is capable of scaling to a large number of noisy time-series. In particular we impose restrictions on the covariance structure so that it is low-rank and sparse, limiting the number of parameters to estimate [cf. 13]. Such sparsity is desirable because it potentially conveys information about the structure of the VAR [24].

A *vector autoregression* model (VAR) consists of a set of K endogenous variables: $\mathbf{y}_t = (y_{1t}, \ldots, y_{kt}, \ldots, y_{Kt})$, $k \in [1, K]$. An spVAR(q) process is defined as

$$\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \ldots + \mathbf{A}_q \mathbf{y}_{t-q} + \mathbf{u}_t$$

where $\mathbf{A}_p$ is a $K \times K$ coefficient matrix, and $\mathbf{u}_t$ is a K-dimensional zero-mean, covariance-stationary innovation process, i.e. $\mathbb{E}[\mathbf{u}_t] = 0$ and $\mathbb{E}[\mathbf{u}_y \mathbf{u}_t^\top] = \Sigma_u$. The *sparse pooling VAR* model (spVAR) further imposes that constraint that the matrices $\mathbf{A}_p$ are low-rank, rank$(\mathbf{A}_p) \leqslant r$ and ℓ_1 sparse [26, 3, 13]. spVAR assumes temporal covariance stationarity; i.e. $\{\mathbf{A}^{(t)}\}_{t \in (1 \ldots T)}$ is constant across the entire history. In practice however, time-series dynamics tend to switch between multiple stationary regimes.[4]

To efficiently estimate spVAR in high dimension we use a variant of Forward-Backward Splitting with forward-looking subgradients [FOBOS; 7]. FOBOS alternates between two steps: (1) an unconstrained gradient descent step, and (2) minimization of the regularization term keeping the solution close to the result of step 1.

4 VAR Forecasting Performance

For VAR forecasting, we are interested in the combinations of data sources that are most predictive of others, focusing in particular on what series can potentially forecast established economic indicators (i.e. the **econometric** data set).

Figure 1 shows individual RMSEs for each response variable broken down by training source, *relative* to training using **econometric**. We find that:

1. The social activity signals **yelp**, **twitter** and **fb.gnh** are better forecasters of **eq.xly** (consumer discretionary sector), **eq.xlk** (technology sector), **fred.icsa** (weekly unemployment claims), and **gdi_us.credit**, **gdi_us.luxury** than the **econometric** data. Furthermore, their RMSE profiles are strongly consistent.
2. Conversely, **econometric** is a significantly better predictor of **eq.xlf** (financials sector), **fred.vixcls** (volatility index) and **gdi.travel** (travel-related searches).

5 Discussion and Future Work

This abstract proposed and evaluated an approach to pooling online activity into fast, approximate measures of macroeconomically relevant variables based on topic modeling and latent-factor vec-

[4] Although we do not investigate it work, note that such regime switching can be implemented efficiently using overlapping group sparsity [19].

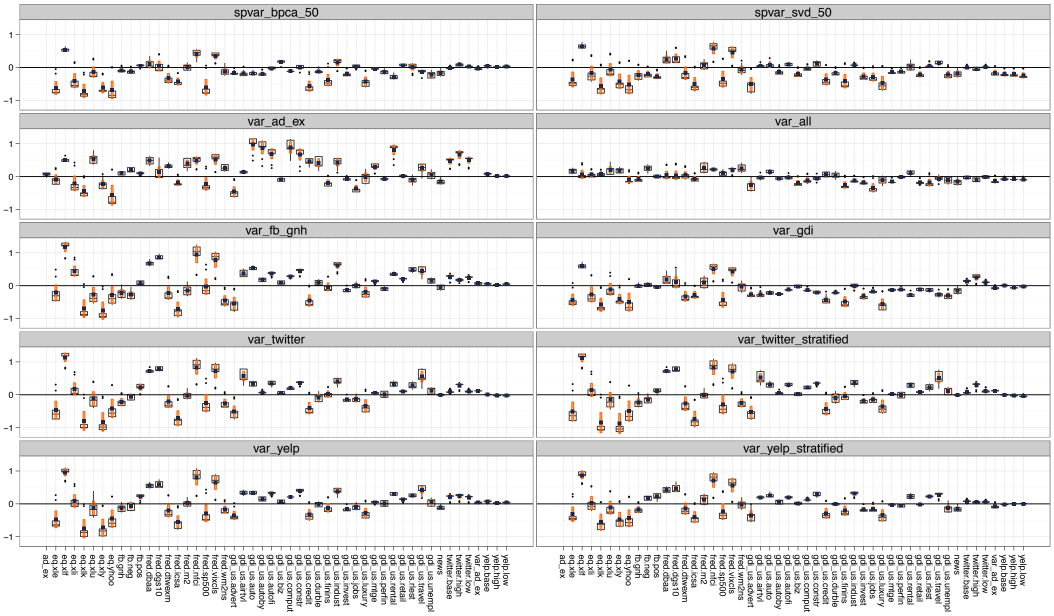

Figure 1: Reduction in forecast RMSE vs. using baseline econometric indicators. Each panel corresponds to a different set of data sources used for forecasting, and relative RMSE performance is broken out over all base data sources. Negative values indicate a reduction in RMSE vs. the baseline **econometric** indicators, while positive values indicate an increase. For example, using the **ad_ex** data sources alone results in a significant reduction in test RMSE when predicting **eq.yhoo** (Yahoo! stock price) and **gdi_us.advert** (advertising related search queries), while the RMSE of predicting aggregate **twitter** activity increases significantly.

tor autoregression. We found that: (1) social network factors such as Yelp reviews, Twitter posts and Facebook sentiment are more predictive of unemployment and consumer discretionary spending than a basket of econometric factors, and (2) this effect is symmetric across various parameter settings of these factors, i.e. Yelp reviews and Twitter posts capture substantially similar variation in economic indicators.

(**"Price Discovery" in Information Flows**) Are there natural analogs to price formation and discovery mechanisms in online activity flows? News events cause reactions and commentary which in turn cause material changes in aggregate sentiment [27]. Models such as spVAR can potentially be used to understand how information flows lead to changes in market prices, and vice-versa. Other approaches include model-free estimators of information flow between markets, such as *transfer entropy* [22], or more traditional microstructure models of price formation [cf. 18].

(**Exogenous Shocks and Temporal Stationarity**) Several data sources exhibit large exogenous shocks; e.g. in the exchange case due to new bidders entering the market or the availability of new targeting information, and in the Twitter case due to spikes in topical content (e.g. "hurricapocalypse"). Switching VAR processes would be better able to account for these regime changes and other sources of heteroskedasticity [10].

References

[1] Google domestic trends index. `www.google.com/finance/domestic_trends`.

[2] S. Asur and B. A. Huberman. Predicting the future with social media. 2010.

[3] F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. *CoRR*, abs/0812.1869, 2008.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

[5] A. Cavallo and R. Rigobon. Billion Prices Project. `bpp.mit.edu`.

[6] H. Choi and H. Varian. Predicting the present with Google trends. Technical report, Google, 2009.

[7] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

[8] B. Efron. Large-Scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[9] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized Dynamic-Factor model: Identification and estimation. *Review of Economics and Statistics*, 82(4):540–554, 2000.

[10] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Sharing features among dynamical systems with beta processes. In *Proc. of NIPS*. MIT Press, 2010.

[11] E. Gilbert and K. Karahalios. Widespread worry and the stock market. In *ICWSM*. The AAAI Press, 2010.

[12] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 2008.

[13] C. Giraud. Low rank multivariate regression. *arXiv:1009.5165v*, 2010.

[14] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, 2010.

[15] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proc. of KDD*, pages 78–87. ACM, 2005.

[16] G. Hale and O. Jórda. Do monetary aggregates help forecast inflation? *FRBSF Economic Letter*, 2007.

[17] J. D. Hamilton. *Time-series analysis*. Princeton Univerity Press, 1994.

[18] J. Hasbrouck. *Empirical Market Microstructure*. Oxford University Press, 2006.

[19] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML*, volume 382. ACM, 2009.

[20] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: an experiment in text regression. In *Proc. of NAACL-HLT 2010*. Association for Computational Linguistics, 2010.

[21] A. D. Kramer. An unobtrusive behavioral model of ”gross national happiness”. In *Proc. of CHI*. ACM, 2010.

[22] R. Marschinski and H. Kantz. Analysing the information flow between financial time series. *The European Physical Journal B - Condensed Matter*, 30(2):275–281, 2002.

[23] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

[24] J. H. Stock and M. W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, 20(2):147–162, 2002.

[25] P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[26] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[27] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Proc. of ICDM*. IEEE, 2010.