MIT Libraries | DSpace@MIT

MIT Open Access Articles

Forecasting with Alternative Data

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Fleder, Michael and Shah, Devavrat. 2020. "Forecasting with Alternative Data."

As Published: https://doi.org/10.1145/3393691.3394187

Publisher: ACM|ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems

Persistent URL: https://hdl.handle.net/1721.1/146169

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Massachusetts Institute of Technology

Forecasting with Alternative Data

Michael Fleder
Massachusetts Institute of Technology
Cambridge, MA, USA
mfleder@mit.edu

Devavrat Shah
Massachusetts Institute of Technology
Cambridge, MA, USA
devavrat@mit.edu

ABSTRACT

We consider the problem of forecasting fine-grained company financials, such as daily revenue, from two input types: noisy proxy signals a la alternative data (e.g. credit card transactions) and sparse ground-truth observations (e.g. quarterly earnings reports). We utilize a classical linear systems model to capture both the evolution of the hidden or latent state (e.g. daily revenue), as well as the proxy signal (e.g. credit cards transactions). The linear system model is particularly well suited here as data is extremely sparse (4 quarterly reports per year). In classical system identification, where the central theme is to learn parameters for such linear systems, unbiased and consistent estimation of parameters is not feasible: the likelihood is non-convex; and worse, the global optimum for maximum likelihood estimation is often non-unique.

As the main contribution of this work, we provide a simple, consistent estimator of all parameters for the linear system model of interest; in addition the estimation is unbiased for some of the parameters. In effect, the additional sparse observations of aggregate hidden state (e.g. quarterly reports) enable system identification in our setup that is not feasible in general. For estimating and forecasting hidden state (actual earnings) using the noisy observations (daily credit card transactions), we utilize the learned linear model along with a natural adaptation of classical Kalman filtering (or Belief Propagation). This leads to optimal inference with respect to mean-squared error. Analytically, we argue that even though the underlying linear system may be "unstable," "uncontrollable," or "undetectable" in the classical setting, our setup and inference algorithm allow for estimation of hidden state with bounded error. Further, the estimation error of the algorithm monotonically decreases as the frequency of the sparse observations increases. This, seemingly intuitive insight contradicts the word on the Street, cf. [7]. Finally, we utilize our framework to estimate quarterly earnings of 34 public companies using credit card transaction data. Our data-driven method convincingly outperforms the Wall Street consensus (analyst) estimates even though our method uses only credit card data as input, while the Wall Street consensus is based on various data sources including experts' input.

KEYWORDS

Forecasting; Alternative Data; Linear Systems; Time Series; Finance; Consumer Credit Card Transactions

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGMETRICS '20 Abstracts, June 8–12, 2020, Boston, MA, USA
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7985-4/20/06.
https://doi.org/10.1145/3393691.3394187

ACM Reference Format:
Michael Fleder and Devavrat Shah. 2020. Forecasting with Alternative Data. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '20 Abstracts), June 8–12, 2020, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3393691.3394187

1 INTRODUCTION

In recent years there has been a proliferation of alternative financial datasets ("alt data") that function as side-channel, or proxy information for company financials [8]. For example, alt data sets consisting of consumer credit card transactions can be used to estimate daily consumer spending on Uber and Lyft [4] or at McDonald's [6]. Ground-truth for these numbers is almost never disclosed on a daily basis, but aggregates (e.g. total revenue) are reported at lower frequency - typically quarterly. Interest in alt data has grown significantly (See Figure 1) primarily because of alt data's promise to help estimate company financials at higher frequency than quarterly reports. However, combining alt data with lower-frequency ground truth (e.g. quarterly reports), for accurate, frequent estimation and forecasting of hidden company financials, remains challenging; and approaches are often highly dataset specific, cf. [1].

The primary goal of this work is to develop a method for accurately forecasting company financials by combining noisy, high-frequency alt data and lower-frequency aggregates (like quarterly reports). In addition, we require our forecasts to update frequently - with every new data point. The flip side of this forecasting exercise is to understand the implications of requiring companies to disclose ground truth in quarterly reports. We investigate how much is really revealed about higher-frequency dynamics through quarterly reports.

1.1 Contributions

We make three contributions. First, the main contribution of this work is providing a systematic approach for tracking company financials at high-frequency, where we combine low-frequency ground-truth aggregates with high-frequency, noisy proxy data. We utilize a variation of the classical linear dynamical systems (LDS) model with hidden state and noisy observations with Gaussian noise. Compared to the classical setting, for example that considered in [3], we are also given infrequent observations of aggregate hidden state. Our goal is the same: to develop an estimation algorithm for the hidden state. We show that our model is effective with sparse observations - for example receiving alternative data weekly but aggregate observations every 3 months.

Our second contribution is solving the two problems required for utilizing this LDS model. First, we solve the problem of learning the model parameters, referred to as "system identification" in the classical literature. We show that the inclusion of infrequent aggregate

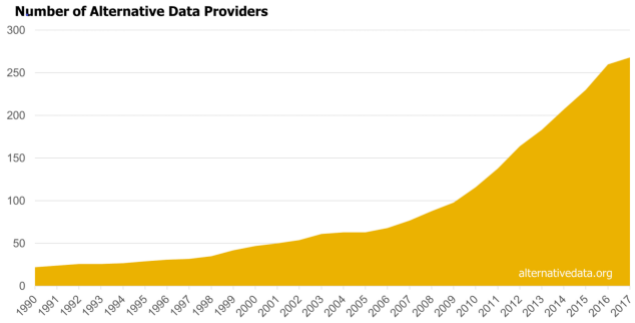

Figure 1: Alternative data providers by year [2]

Metric	LDS	Benchmark
RMSE	2.7	3.2
Median Abs Error	1.2	1.3
Wins (Total Quarters)	175	131
Win Percent	57.2%	42.8%

Table 1: Linear dynamical system (LDS) versus Wall Street consensus benchmark. We learn a separate LDS model for each train/test split of each company's data. We use leave-one-out cross-validation and report the resulting test set performance. Test performance is aggregated across all 34 companies and 306 quarters. By "win" we mean a test quarter for which the LDS outperforms the benchmark. The in-general, low-percent error of the benchmark indicates the difficulty of the forecasting task. The LDS win percentage is statistically significant (see Section 4.2.3 in the paper).

Quarter	LDS Abs Error(%)	Benchmark Abs Error(%)
1	22.4%	16.2%
2	9.1%	1.5%
3	**4.4%**	5.8%
4	**5.6%**	6.1%
5	18.4%	11.8%
6	**3.0%**	18.2%
7	4.6%	2.2%
8	**8.3%**	16.7%
9	**5.9%**	10.2%

Table 2: Sample results for one company over 9 quarters. In this instance, the LDS model outperforms the benchmark in 5/9 quarters. We highlight in green (and bold) the quarters for which the LDS model outperforms the benchmark. Overall, the mean-absolute percent error for LDS here is 9.1% versus 9.9% for the benchmark.

state observations allows us to devise consistent estimators for all parameters, and provide a finite sample analysis of the resulting error. And for some of the parameters, the estimation is unbiased. This is surprising, because in the classical setup, system identification for LDS suffers from non-uniqueness: multiple sets of parameters give rise to identical likelihood values ([5] page 387); and furthermore, the likelihood is non-convex, making optimization challenging (see Figure 3 in the paper). In contrast, our algorithms are consistent and computationally efficient: given k observations of the latent state, a latent state of dimension n, and observations of dimension m, the running time of our system identification algorithms depends only on multiplication of matrices of size $(k \times n), (k \times m)$ and inversion of $n \times n$ matrices, for which there are efficient methods.

The next part of this contribution, of utilizing a LDS in a sparse data setting, is providing and analyzing an optimal inference algorithm. In the classical setting, Kalman Filtering, provides the optimal estimation procedure in terms of minimizing mean squared error. In our modified setup, this no longer holds. We develop such a method and provide an optimal estimation algorithm. We show that if ground-truth aggregate information is available with any non-zero expected frequency, then tracking estimation error remains bounded - even if the dynamics are unstable, uncontrollable, or undetectable. Furthermore, we show that the tracking estimation error decreases monotonically as ground-truth information becomes increasingly available - which is intuitively pleasing. This directly contradicts the claim General Motors Co. (GM) made in April 2018, when GM switched from monthly to quarterly reporting for its U.S. vehicle sales; GM stated that monthly sales are not useful for investors [7]. Per our analysis, having monthly versus quarterly earnings reports improves estimation of company financials; thus GM's claim is arguably incorrect.

Our third contribution is putting our model into practice. We start by empirically validating the theorems using synthetic data. Next, we apply our end-to-end identification and tracking algorithms to an alternative data set of real credit card transactions obtained from a hedge fund. The data set consists of typically weekly or biweekly summaries of unknown fractions of consumer spending at 34 public companies. The prediction task is to forecast weekly revenue (and hence quarterly earnings) at each of the companies using both the credit card data along with historical, public, quarterly disclosures of revenue. Our method outperforms a standard benchmark of Wall Street consensus estimates, beating the consensus on 57.2% of quarterly predictions as well as outperforming with respect to the mean-squared error (See Tables 1 and 2). The model performance is significant, because we do not make use of any additional information or expert input that may have been available as input for other financial analysts' estimates.

REFERENCES

[1] Eagle Alpha. 2018. Eagle Alpha Alternative Data Use Cases. https://eaglealpha.com/eagle-alphas-alternative-data-use-cases. Accessed: 2018-05-10.
[2] AlternativeData.org. 2018. Alternative Data by the Numbers. https://alternativedata.org/resources/alternative-data-by-the-numbers. Accessed: 2018-05-17.
[3] Dimitri P Bertsekas. 1995. *Dynamic programming and optimal control*. Vol. 1. Athena scientific, Belmont, MA.
[4] Amir Efrati. 2018. U.S. Slowdown at Uber and Lyft. https://www.theinformation.com/articles/u-s-slowdown-at-uber-and-lyft. Accessed: 2018-10-25.
[5] James Douglas Hamilton. 1994. *Time series analysis*. Princeton Univ. Press, Princeton, NJ.
[6] Bradley Hope. 2015. Provider of Personal Finance Tools Tracks Bank Cards Sells Data to Investors. https://www.wsj.com/articles/provider-of-personal-finance-tools-tracks-bank-cards-sells-data-to-investors-1438914620. Accessed: 2018-05-10.
[7] Joseph White. 2018. GM to drop monthly U.S. vehicle sale reports. https://www.reuters.com/article/us-usa-autos-gm/gm-to-drop-monthly-u-s-vehicle-sale-reports-idUSKCN1HA0C9. Accessed: 2018-05-07.
[8] Robin Wigglesworth. 2018. Asset management's fight for alternative data analysts heats up. https://www.ft.com/content/2f454550-02c8-11e8-9650-9c0ad2d7c5b5. Accessed: 2018-05-07.