I Know What You Bought At Chipotle for $9.81 by Solving A Linear Inverse Problem

I Know What You Bought At Chipotle for $9.81 by Solving A Linear Inverse Problem

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

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

ABSTRACT

We consider the question of identifying which set of products are purchased and at what prices in a given transaction by observing only the total amount spent in the transaction, and nothing more. The ability to solve such an inverse problem can lead to refined information about consumer spending by simply observing anonymized credit card transactions data. Indeed, when considered in isolation, it is impossible to identify the products purchased and their prices from a given transaction based on just the transaction total. However, given a large number of transactions, there may be a hope.

As the main contribution of this work, we provide a robust estimation algorithm for decomposing transaction totals into the underlying, individual product(s) purchased by utilizing a large corpus of transactions. Our method recovers a (product prices) vector $p \in \mathbb{R}^N_{>0}$ of unknown dimension (number of products) N as well as matrix $A \in \mathbb{Z}^{M\times N}_{\geq 0}$ simply from M observations (transaction totals) $y \in \mathbb{R}^M_{>0}$ such that $y = Ap + \eta$ with $\eta \in \mathbb{R}^M$ representing noise (taxes, discounts, etc.). We formally establish that our algorithm identifies N, A precisely and p approximately, as long as each product is purchased individually at least once, i.e. $M \geq N$ and A has rank N. Computationally, the algorithm runs in polynomial time (with respect to problem parameters), and thus we provide a computationally efficient and statistically robust method for solving such inverse problems.

We apply the algorithm to a large corpus of anonymized consumer credit card transactions in the period 2016-2019, with data obtained from a commercial data vendor. The transactions are associated with spending at Apple, Chipotle, Netflix, and Spotify. From just transactions data, our algorithm identifies (i) key price points (without access to the listed prices), (ii) products purchased within a transaction, (iii) product launches, and (iv) evidence of a new 'secret' product from Netflix - rumored to be in limited release.

KEYWORDS

Blind Compressed Sensing; Alternative Data; Finance; Consumer Credit Card Transactions

SIGMETRICS '21 Abstracts, June 14–18, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8072-0/21/06.
https://doi.org/10.1145/3410220.3456273

ACM Reference Format:
Michael Fleder and Devavrat Shah. 2021. I Know What You Bought At Chipotle for $9.81 by Solving A Linear Inverse Problem . In *Abstract Proceedings of the 2021 ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '21 Abstracts), June 14–18, 2021, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3410220.3456273

1 INTRODUCTION

Tracking granular consumer spending is of great interest to advertisers, hedge funds [5] banks [1], and others studying the retail economy. Advertisers like Google purchase transactions data to measure in-store retail sales. Similarly, retailers track competitors through such data [7]. And hedge funds utilize transactions data in tracking public companies and informing investment and risk decisions [4]. The prevalence of transactions, primarily via credit and debit cards, has led to anonymized consumer transactions data becoming widely available from a variety of commercial data vendors [7].

Although transactions data details how much consumers spend in *total* at a given vendor; the data does not reveal *what* consumers are buying. For example, a single transaction total at an Apple store for $182.91 does not provide information on which products were purchased or at what prices.

In this work, we are interested in inferring the product(s) a consumer purchases, given knowledge of only the transaction total (a single number). Inferring the products that constitute transaction totals can lead to a wealth of insights. For example, such inference enables estimation of iPhone unit sales at Apple or guacamole sales at Chipotle. Such unit-sales estimates would enable demand estimation for elements upstream the supply chain, e.g. chipmakers for iPhones or food growers for Chipotle.

1.1 Problem Statement

Formally, we have access to M separate, but not necessarily distinct, transaction totals $y \in \mathbb{R}^M_{>0}$ associated with a given company. We assume the company has a finite number of N products with associated prices $p \in \mathbb{R}^N_{>0}$. We treat both N and p as unknown: that is, an unknown number of products with unknown prices. Each transaction total corresponds to the summation of prices for the products purchased. In addition, we include an additive noise term η to account for price variations due to discounts, promotions, taxes, etc. Therefore, we have

$$y = Ap + \eta. \tag{1}$$

Company	Number of Transactions	Our Findings
Netflix	2.6M	Secret Product Ultra HD ($17.08)
Spotify	387K	Product Launch Detected ($12.99)
Apple	197K	iPhone XS Sales Volume
Chipotle	133K	Decomposition Err MAPE < 2%

Table 1: Dataset summary of anonymized transactions.

Result	A	p	N	Noise	Requirements
Ordinary Least Sq.	Known (Full Rank [6])	Unknown	Known	Yes	$M \gg N$
Compressed Sensing	Known (RIP [2, 3])	Unknown	Known	Yes	$M \gg \|\|p\|\|_0$
This Work	Unknown (Signature)	Unknown	Unknown	Yes	$M \gg \|\|p\|\|_0$

Table 2: Succinct comparison of ours with relevant prior works.

where the unknown matrix $A \in \mathbb{Z}_{\geq 0}^{M \times N}$ represents the product decomposition for each transaction. The goal is to identify the number of products N, their associated prices p, and the decomposition of transactions A; all from observation of y only.

1.2 Contributions

A novel inference algorithm. As the main contribution of this work, we develop a simple, iterative and computationally efficient algorithm for inferring N, A and p from y. The algorithm provably recovers N, A precisely, and p approximately, with approximation error dependent on the ℓ_∞-norm of noise η. Our algorithm succeeds if A satisfies a "Signature Condition," which requires that every product is purchased individually at least once. This requires that $M \geq \|p\|_0 = N$. It also guarantees that A has full column rank as required in traditional linear regression (or ordinary least squares). However, the Signature Condition does not require restricted iso-perimetry-like (RIP) conditions which are common in literature on compressed sensing, cf. [2, 3]. We emphasize that A is treated as unknown, whereas compressed sensing literature commonly assumes A is known (see Table 2).

The algorithm recovers prices that are distinct and not multiples of each other. Indeed, no algorithm can distinguish if one or more products are sold at the same price (or as different integral multiples of the same value) without additional side information. Therefore, we require such a condition to hold for proving that we accurately recover N, A, p from just y.

The algorithm requires solving M exact subset-sum problems which is known to be NP-hard. However we prove that it admits a fully polynomial time (in problem parameters) implementation based on approximate subset-sum that achieves similar performance guarantees.

In summary, we provide a simple, iterative and computationally efficient algorithm for solving the system of linear equations *without* knowledge of A, p or N – unlike any prior works (see Table 2). This might be of interest in its own right.

Empirical validation. We apply the algorithm to anonymized consumer transactions data[1] based on credit and debit card purchases at Netflix, Spotify, Apple and Chipotle (see Table 1). We measure performance of the algorithm in terms of recovering (a) the number of key products N, (b) their corresponding prices p, and (c) the decomposition (A) of transaction totals into products purchased. In addition, we discuss the implications of accurate inference in terms of detecting product launches and hidden / non-advertised product offerings.

For (a) and (b), it is easy to verify full recovery (or not) of all products for Netflix, since Netflix has few product offerings. We recover all the published offerings by Netflix. In addition, we identify two additional 'hidden' offerings which seem to agree with limited-release products. Our method has a median error in the recovered prices of less than 0.2% (or, less than 4 cents). Accurate recovery of the number of products and their prices implies that that each transaction total is accurately modeled in terms of identifying the product(s) purchased, i.e. performance in terms of (c). For Chipotle, which offers a larger and more complex set of offerings, we reconstruct transaction totals within MAPE of 1.2% using 12 key product prices only!

Our findings enable a plethora of insights into time-varying company product catalogues. For example, at Spotify and Apple, we utilize our methods to automatically detect product launches at both new and existing price points – all from anonymized transaction totals only.

Collectively, these experiments verify that our method is able to recover product prices as well as decompose transaction totals accurately across different types of businesses: from Netflix and Spotify with few offerings, to Chipotle and Apple with extremely complex product offerings. Indeed, our method is likely to have more impactful consequences such as estimating sales volume by price range.

[1] We utilized anonymized debit and credit card transactions data provided by alternative data company Second Measure [7] for the purpose of conducting this work. Table 1 provides dataset details.

REFERENCES

[1] Florentin Butaru, QingQing Chen, Brian Clark, Sanmay Das, Andrew W Lo, and Akhtar Siddique. Risk and risk management in the credit card industry. Working Paper 21305, National Bureau of Economic Research, June 2015.

[2] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[3] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[4] Michael Fleder and Devavrat Shah. Forecasting with alternative data. In *Abstracts of the 2020 SIGMETRICS/Performance Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '20, page 23–24, New York, NY, USA, 2020. Association for Computing Machinery.

[5] IO&C. The big trends in data reshaping financial industry. https://ioandc.com/the-big-trends-in-data-reshaping-financial-industry, April 2019. Accessed: 2019-04-07.

[6] Tze Leung Lai, Ching Zong Wei, et al. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

[7] Second Measure. Data points. https://secondmeasure.com/datapoints. Accessed: 2019-05-19.