Censored Demand Estimation in Retail

Muhammad J Amjad

Massachusetts Institute of Technology
Cambridge, MA, USA
mamjad@mit.edu

Devavrat Shah

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

ABSTRACT

In this paper, the question of interest is estimating *true* demand of a product at a given store location and time period in the retail environment based on a *single* noisy and potentially censored observation. To address this question, we introduce a framework to make inference from multiple time series. Somewhat surprisingly, we establish that the algorithm introduced for the purpose of “matrix completion” can be used to solve the relevant inference problem. Specifically, using the Universal Singular Value Thresholding (USVT) algorithm [2], we show that our estimator is consistent: the average mean squared error of the estimated average demand with respect to the true average demand goes to 0 as the number of store locations and time intervals increase to ∞ . We establish naturally appealing properties of the resulting estimator both analytically as well as through a sequence of instructive simulations. Using a real dataset in retail (Walmart), we argue for the practical relevance of our approach.

CCS CONCEPTS

• **General and reference** → **Estimation**; • **Computing methodologies** → *Latent variable models*; • **Theory of computation** → Sample complexity and generalization bounds;

KEYWORDS

Singular value thresholding; Matrix completion; Censored demand; Estimation

ACM Reference Format:

Muhammad J Amjad and Devavrat Shah. 2018. Censored Demand Estimation in Retail. In *SIGMETRICS '18 Abstracts: ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems Abstracts, June 18–22, 2018, Irvine, CA, USA*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3219617.3219624>

Background.

Consider the problem of estimating the daily or weekly rate at which umbrellas are sold at a specific location, say at the Walmart store in Bentonville, Arkansas. To do so, we have *one* sample per time unit across several stores, e.g. 4 and 3 umbrellas were sold in the past two weeks at store A, 6 and 5 were sold at store B and so on. The problem is challenging because the observations can be noisy, incomplete and censored. The noise is due to random errors in measurement or record-keeping (e.g. mismatch in inventory records

and physical stocks, transaction errors). The data might also be incomplete due to missed reporting or aggregations for some days or weeks. Importantly, the data is censored because the store might have stocked only 4 umbrellas during the past week and, hence, observed 4 sales but there was no information to account for any customers who might have wished to purchase an umbrella but could not do so due to the stock-out. In this paper we provide a simple inference algorithm to estimate the time varying demand rate from effectively a single noisy, incomplete and censored observation across multiple locations. The key enabler for this is the latent variable model which allows us to utilize information across a number of stores to synthetically create “multiple observations” for a given time unit and location to better estimate the time-varying demand rate.

Setup and Contributions

While the problem of estimating *true* demand given censored sales data has been studied extensively, existing models have proven to be unsatisfactory in terms of faithfully capturing reality. Specifically, the problem of inferring time varying demand based on censored information that is sparse, a focus of this work, has remained unresolved. As the key contribution of this work, we provide a model to capture this scenario through the “latent variable model”. Through the lens of this latent variable model, we are able to apply the rich literature on “matrix estimation” to enable effective resolution of the problem at hand.

We consider a setting where a retailer has censored sales data for a product or group of products across N store locations and T time periods. Without loss of generality, we shall assume that $N \leq T$. Let *true* demand at each location and for each time period be modeled as an independent random variable with Poisson distribution. Specifically, let Y_{ij} denote the *true* demand at store $1 \leq i \leq N$ at time $1 \leq j \leq T$ with $\lambda_{ij} = \mathbb{E}[Y_{ij}]$ being the mean demand. In matrix form, let $Y = [Y_{ij}]_{i \leq N, j \leq T}$ and $\Lambda = [\lambda_{ij}]_{i \leq N, j \leq T}$.

Let C_{ij} be the quantity of stock (or inventory) at store $i \leq N$ during time interval $j \leq T$. Therefore, the number of sales, $X_{ij} = \min(Y_{ij}, C_{ij})$. That is, X_{ij} represents the *censored* demand at store i at time j . We let $m_{ij} = \mathbb{E}[X_{ij}]$. In matrix form, let $C = [C_{ij}]_{i \leq N, j \leq T}$, $X = [X_{ij}]_{i \leq N, j \leq T}$ and $M = [m_{ij}]_{i \leq N, j \leq T}$.

To model the situation where some stores might not have reported any information at various time periods due to supply chain issues, information mismanagement, etc., we consider a setup where each X_{ij} is observed with probability $p \in (0, 1]$ and not observed with probability $1 - p$, independently. Let X^p denote this *partially observed* matrix of censored demand matrix X . The goal is to estimate Λ from X^p as accurately as possible.

To that end, if there is *no structure* in Λ , there is no hope to obtain any meaningful estimate of Λ from X^p . For example, let $p = 1$, let C_{ij} be very large (say, ∞) for all $i \leq N, j \leq T$, and let each

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 '18 Abstracts, June 18–22, 2018, Irvine, CA, USA

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5846-0/18/06.

<https://doi.org/10.1145/3219617.3219624>

λ_{ij} be arbitrarily chosen. Then effectively we are observing one sample each of $N \times T$ Poisson random variables that have nothing to do with each other. Equivalently, for a given i, j , we are trying to estimate mean λ_{ij} of a Poisson variable from one sample. Naturally, that is a futile exercise.

Therefore, to obtain a meaningful estimate, it is essential to impose structure. In the context of retail, it makes sense that the average demand at store $i \leq N$ at time interval $j \leq T$ depends on the store as well as the time period itself. Formally, let $\lambda_{ij} = h(\theta_i, \rho_j)$ where θ_i and ρ_j are *latent* or *hidden* features associated with the store i and time j ; and h is an arbitrary Lipschitz continuous function. This is in contrast to the standard assumptions in literature where the latent matrix is assumed to have a low-rank structure. The Lipschitz structure leads to a more generic model and seems to have enough expressive power to capture reality well.

As the main result of this work, we provide an estimation algorithm for Λ using X^P such that the expected mean squared error (MSE), with respect to Λ , in the estimate $\hat{\Lambda}$ goes to 0 as $N \rightarrow \infty$ as long as $p \gg N^{-\frac{2}{d+2}}$ where d is the dimension of the latent feature space. We also discuss that the imposition of a more realistic Lipschitz structure instead of the low-rank assumption comes with the cost of a slower decay of the MSE, as N increases.

Our estimation algorithm is a two step procedure: in the first step, it produces an estimate $\hat{M}$ of M from X^P ; in the second, it produces as estimate $\hat{\Lambda}$ of Λ using $\hat{M}$ and knowledge of C . See below for the precise details of the algorithm.

To produce $\hat{M}$ using X^P , we utilize the Universal Singular Value Thresholding (USVT) algorithm by Chatterjee [2]. Effectively, the algorithm computes the singular value decomposition (SVD) of X^P ; truncates the decomposition by keeping only top few singular vectors / values and multiplies it by an appropriate parameter. The choice of the number of top singular vectors / values to retain is done *universally* based only on p and dimension of the matrix, hence, *universal* singular value thresholding. To bound the expected $MSE(\hat{M})$, with respect to M , under the setup described earlier, we provide a minor modification of the result established in [2]

To produce $\hat{\Lambda}$ from $\hat{M}$ using knowledge of C , we use analytic properties of the (truncated) Poisson distribution along with a natural “bisection” algorithm. Using elementary calculations, we establish that the expected $MSE(\hat{\Lambda})$, with respect to Λ , is within constant factor of the expected $MSE(\hat{M})$, with respect to M ; the constant primarily depends on C . This constant factor gets close to 1 as the entries in C increase; it becomes larger as entries in C decrease. Intuitively, this makes sense – as the entries of C increase, the effect of censoring disappears and, hence, M becomes closer to Λ , and vice versa.

The Algorithm

We are given partial observations of the censored demand matrix, X^P . We wish to produce an estimate $\hat{\Lambda}$ of true average demand Λ . We propose to do so in two steps: (1) Obtain an estimate of the average censored demand, i.e. $\hat{M}$ of $M = \mathbb{E}[X]$, and (2) extrapolate $\hat{M}$ to obtain $\hat{\Lambda}$ using the knowledge of capacity matrix C .

Step 1. Obtaining $\hat{M}$. We apply the Universal Singular Value Thresholding (USVT) of [2] to X^P to obtain $\hat{M}$. For completeness, we describe the USVT algorithm [2]:

- (1) Define $b = \max_{i,j} z_{ij}$ and $a = \min_{i,j} z_{ij}$.
- (2) $z_{ij} \leftarrow \frac{z_{ij} - (a+b)/2}{(b-a)/2}$. Now, $|z_{ij}| \leq 1, \forall i, j$.
- (3) Define matrix $Z = [z_{ij}]_{i \leq N, j \leq T}$ with

$$z_{ij} = \begin{cases} X_{ij} & \text{if it is observed in } X^P \\ 0 & \text{otherwise.} \end{cases}$$

- (4) Let $Z = \sum_{i=1}^N s_i u_i v_i^T$ be the singular value decomposition of Z .
- (5) Let $\hat{p}$ be fraction of the NT entries observed in X^P , i.e. empirical estimation of p based on number of entries observed.
- (6) Let $S = \{i : s_i \geq 2.02\sqrt{T\hat{p}}\}$.
- (7) Define $W = \frac{1}{\hat{p}} \sum_{i \in S} s_i u_i v_i^T$.
- (8) Let w_{ij} be the (i, j) th element of W . Define

$$\hat{m}_{ij} = \begin{cases} -1 & \text{if } w_{ij} < -1 \\ 1 & \text{if } w_{ij} > 1 \\ w_{ij} & \text{otherwise.} \end{cases} \quad (1)$$

- (9) Scale back to the original range:

$$\hat{m}_{ij} \leftarrow (a+b)/2 + \hat{m}_{ij}(b-a)/2. \quad (2)$$

Step 2. Obtaining $\hat{\Lambda}$. We have access to $\hat{M}$, the estimate of M where the (i, j) th element $\hat{m}_{ij}$ of $\hat{M}$ is an estimate of $m_{ij} = \mathbb{E}[X_{ij}]$, the (i, j) th element of M , which is the average of truncated Poisson random variable with mean λ_{ij} , truncated at C_{ij} . From $\hat{M}$, we want to produce $\hat{\Lambda}$, an estimate of Λ , using knowledge of C , which is known.

A simple iterative algorithm (this is also known as the *Bisection* algorithm in literature) to find λ_{ij} is as follows:

- (1) Initialize $\lambda_{ij}^{UB} = \infty, \lambda_{ij}^{LB} = 0$ and $\lambda_{ij}^1 = 1$.
- (2) In iteration $k \geq 1$, let $m_{ij}^k = f(\lambda_{ij}^k, C_{ij})$. If $m_{ij}^k > m_{ij}$ then update $\lambda_{ij}^{UB} = \lambda_{ij}^k$. If $m_{ij}^k < m_{ij}$, update $\lambda_{ij}^{LB} = \lambda_{ij}^k$. And,

$$\lambda_{ij}^{k+1} = \begin{cases} \frac{1}{2}(\lambda_{ij}^{UB} + \lambda_{ij}^{LB}), & \text{if } \lambda_{ij}^{UB} < \infty \\ 2\lambda_{ij}^{LB}, & \text{if } \lambda_{ij}^{UB} = \infty. \end{cases} \quad (3)$$

- (3) Stop iterating when $|\lambda_{ij}^{UB} - \lambda_{ij}^{LB}|$ is small enough and declare estimate of $\lambda_{ij} = \frac{1}{2}(\lambda_{ij}^{UB} + \lambda_{ij}^{LB})$.

In reality, we do not know m_{ij} , but we know estimate for it, $\hat{m}_{ij}$. Therefore, we use $\hat{m}_{ij}$ in place of m_{ij} in the above algorithm. We denote the resulting estimation of Λ by $\hat{\Lambda}$.

Summary of Experiments

Synthetic Data. While our theorems provide useful bounds, we conduct extensive synthetic experiments to understand the finer performance dependency of the estimation algorithm, not fully explained by our theoretical results. As mentioned earlier, our key result is the bound on $MSE(\hat{\Lambda})$ in terms of $MSE(\hat{M})$. To understand the behavior of this constant factor in the bound as a function of censoring, we vary the degree of censoring and find that as censoring decreases (equivalently, entries of C increase) the bound

Figure 1: Effect of increasing the size of the matrix, X^P , on the $RMSE(\hat{\Lambda})$. The different plots represent different levels of censoring. $J = 5$. The three levels of censoring are: little ($C = 10$), mild ($C = 5$) and significant ($C = 2$).

Figure 2: Effect of increasing the size of the matrix, X^P , on the Ratio: $\frac{RMSE(\hat{\Lambda})}{RMSE(\hat{M})}$. The different plots represent different levels of censoring. $J = 5$. The three levels of censoring are: little ($C = 10$), mild ($C = 5$) and significant ($C = 2$).

decreases and vice versa. See Figures 1 and 2. Note that somewhat counter-intuitively, as the entries in C increase, the $MSE(\hat{M})$ increases. This behavior can be explained by realizing that as entries in C increase, the “support” of random variables Y_{ij} increases. We also note that the bound remains unaffected by the size of the matrix, even though the $MSE(\hat{\Lambda})$ and $MSE(\hat{M})$ themselves decrease.

Walmart Data. We used sales data published by Walmart on Kaggle [1] to conduct our experiments with the hope of understanding the applicability as well as impact of our results in a practical setting. This dataset contains sales data for several departments across 45 store locations and 143 weeks (time periods). Clearly, we do not have the knowledge of the ground truth in terms of the underlying “generative model” like in the case of synthetic data. Further, we do not have access to inventory information. We apply our method based on the model described earlier.

To begin with, we wanted to find evidence in the data about validity of structure across stores and time periods as considered in this paper. If there is a meaningful structure that our algorithm exploits, then we should find as the fraction, p , of observed data increases, we should be able to reconstruct missing information with higher accuracy. And we do find that (see Figure 3).

Next, we wish to verify whether our model assumption the each store and time period’s demand can be modeled as independent (but different) Poisson random variable makes sense. To that end, we conduct the following experiment: for each store and time, we find the mean parameter using our method. For Poisson, the mean parameter tells us about the variance. If there is independence, then

Figure 3: For the Walmart sales data across 45 stores and 143 weeks. Department = 79. The plot shows $RMSE(\hat{\Lambda})$ vs observation probability, p . The RMSE is obtained between the estimated $\hat{\lambda}_{ij}$ and the original observations Y . We assume little to no-censoring. The plot is showing RMSE values for the entire matrix $\hat{\Lambda}$ and also for only those values that were hidden (due to our choice of p).

Figure 4: For the Walmart sales data across 45 stores and 143 weeks. Department = 79. The histogram on the left is generated by random sampling from the random variable $\mathbb{W}$. The matrices are sampled at random with an independent probability of selection, i.e. $p < 1$. On the the right is a QQ-plot of the distribution of $\mathbb{W}$ against a Normal distribution with mean being the mean of the samples of $\mathbb{W}$ and standard deviation of the samples of $\mathbb{W}$.

we can determine the overall variance. Interestingly enough, this “model based” variance estimation matches the overall empirical variance (see Figure 4). This suggests that data is not contradicting our model assumption. This is important because while our methodology extends to other distributions, if the “true” distribution is different from the one assumed, an error will be introduced quantifying which is not the a consideration in this work.

For the Walmart case study, it is important to note that the estimated censored demand is non-trivially different from the observation suggesting that there is “learning” to be done from the data. The average of the estimated means are noticeably smaller than the empirical average suggesting that there is non-trivial censoring happening in the data. Of course, we could have explicitly verified this if we had access to the inventory information. Finally, it is easy to see that the $\hat{M}$ is a lower bound to Λ ; that is, $\hat{M}$ is an estimation of a lower bound of true demand.

ACKNOWLEDGMENTS

This work is supported in parts by NSF under grants CMMI-1462158, CMMI-1634259, by DARPA under grant W911NF-16-1-0551.

REFERENCES

- [1] 2014. <https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting>
- [2] Sourav Chatterjee. 2015. Matrix estimation by Universal Singular Value Thresholding. *Ann. Statist.* 43 (2015), 177–214.