Censored Demand Estimation in Retail

MUHAMMAD J AMJAD, Massachusetts Institute of Technology, USA
DEVAVRAT SHAH, Massachusetts Institute of Technology, USA

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 [7], 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;

Additional Key Words and Phrases: Singular value thresholding; Matrix completion; Censored demand; Estimation

ACM Reference Format:
Muhammad J Amjad and Devavrat Shah. 2017. Censored Demand Estimation in Retail. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 2, Article 31 (December 2017), 28 pages. https://doi.org/10.1145/3154489

1 INTRODUCTION

1.1 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. This is in contrast to online (web) portals which tend to have good estimates of missed demand due to their ability to

Authors' addresses: Muhammad J Amjad, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, USA, 02139, mamjad@mit.edu; Devavrat Shah, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, USA, 02139, devavrat@mit.edu.

2476-1249/2017/12-ART31 $$15.00
https://doi.org/10.1145/3154489

view customer arrival during stock-outs. Additionally, note that the *true* (uncensored) demand is likely to change from week to week, further complicating our problem of estimating it.

It has successfully been shown that ignoring censoring effects will result in demand estimates that are biased lower than the *true* value [27]. Furthermore, as one can intuitively expect, the lack of a complete picture, i.e. censoring, can have a costly impact on inventory planning exercises [8]. In [8], it is suggested that the impact of a lack of complete visibility can be overcome using "intelligent analytics". In this paper, as an example of "intelligent analytics", 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.

1.2 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[1]. 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 $\boldsymbol{Y} = [Y_{ij}]_{i \leq N, j \leq T}$ and $\boldsymbol{\Lambda} = [\lambda_{ij}]_{i \leq N, j \leq T}$.

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 $\boldsymbol{C} = [C_{ij}]_{i \leq N, j \leq T}$, $\boldsymbol{X} = [X_{ij}]_{i \leq N, j \leq T}$ and $\boldsymbol{M} = [m_{ij}]_{i \leq N, j \leq T}$.

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 $\boldsymbol{X}^p$ denote this *partially observed* matrix of censored demand matrix $\boldsymbol{X}$. The goal is to estimate $\boldsymbol{\Lambda}$ from $\boldsymbol{X}^p$ 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 $\boldsymbol{X}^p$. For example, let $p = 1$, let C_{ij} be very large (say, ∞) for all $i \leq N, j \leq T$, and let each 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

[1] Our methodology will work for other distributions as well, provided that the independence assumption is satisfied. See Appendix A for a similar development with an alternate distribution

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 $\boldsymbol{\Lambda}$ using $\boldsymbol{X}^p$ such that the expected mean squared error (MSE), with respect to $\boldsymbol{\Lambda}$, in the estimate $\hat{\boldsymbol{\Lambda}}$ goes to 0 as $N \longrightarrow \infty$ as long as $p \gg N^{-\frac{2}{d+2}}$ where d is the dimension of the latent feature space. See Theorem 4.3 for precise details. In Section 4.3 we 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 $\hat{\boldsymbol{M}}$ of $\boldsymbol{M}$ from $\boldsymbol{X}^p$; in the second, it produces as estimate $\hat{\boldsymbol{\Lambda}}$ of $\boldsymbol{\Lambda}$ using $\hat{\boldsymbol{M}}$ and knowledge of $\boldsymbol{C}$.

To produce $\hat{\boldsymbol{M}}$ using $\boldsymbol{X}^p$, we utilize the Universal Singular Value Thresholding (USVT) algorithm by Chatterjee [7]. Effectively, the algorithm computes the singular value decomposition (SVD) of $\boldsymbol{X}^p$; 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 $MSE(\hat{\boldsymbol{M}})$, with respect to $\boldsymbol{M}$, under the setup described earlier, we provide a minor modification of the result established in [7] stated through Lemma 5.4 and Theorem 5.5. For completeness, we provide the proof for these results, which are direct adaptions from [7]. In Section 3.1, we discuss the advantages of using the USVT algorithm when compared to a somewhat related algorithm in literature [16].

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

1.3 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 $MSE(\hat{\boldsymbol{\Lambda}})$ in terms of $MSE(\hat{\boldsymbol{M}})$. 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 $\boldsymbol{C}$ increase) the bound decreases and vice versa. However, somewhat counter-intuitively, as the entries in $\boldsymbol{C}$ increase, the $MSE(\hat{\boldsymbol{M}})$ increases. This behavior can be explained by realizing that as entries in $\boldsymbol{C}$ 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 $MSE(\hat{\boldsymbol{\Lambda}})$ and $MSE(\hat{\boldsymbol{M}})$ 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.

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 we can determine the overall variance. Interestingly enough, this "model based" variance estimation matches the overall empirical variance. This suggests that data is not contradicting our model assumption. This is important because while our methodology extends to other distributions, see Appendix A for an example, 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 $\boldsymbol{M}$ is a lower bound to $\boldsymbol{\Lambda}$; that is, $\hat{\boldsymbol{M}}$ is an estimation of a lower bound of true demand.

1.4 Notations

We shall use $\mathbb{R}$ to denote all real values, $\mathbb{R}_+$ to denote strictly positive real values, $\mathbb{Z}$ represents all integers, $\mathbb{Z}_+$ represent strictly positive integers. For any $A \in \mathbb{Z}_+$, $[A]$ represents $\{1, \ldots, A\}$. For an $a \times b$ real-valued matrix $Q = [Q_{ij}]$, its Frobenius norm, denoted by $\|Q\|_F$, is given by $\|Q\|_F = \left(\sum_{i=1}^{a}\sum_{j=1}^{b} Q_{ij}^2\right)^{\frac{1}{2}}$. The nuclear norm of Q, denoted by $\|Q\|_*$, is defined as $\|Q\|_* = \sum_{i=1}^{\min(a,b)} s_i$, where s_i, $1 \leq i \leq \min(a,b)$ are singular values of Q.

Given an $a \times b$ matrix Q, let $\hat{Q}$ be a random matrix that is an estimator of Q. Then the error in this estimator, denoted as average mean squared error, denoted as $MSE(\hat{Q})$, is defined as

$$MSE(\hat{Q}) = \frac{1}{ab}\mathbb{E}\left[\|Q - \hat{Q}\|_F^2\right]. \tag{1}$$

The root mean squared error, denoted as $RMSE(\hat{Q})$ is simply defined as square-root of $MSE(\hat{Q})$, that is, $RMSE(\hat{Q}) = \sqrt{MSE(\hat{Q})}$.

1.5 Organization

The rest of this work is organized as follows: we review the relevant domains of literature in Section 2. We describe the estimation algorithm in Section 3. Subsequently, we present the main result (Section 4) and the associated proofs (Section 5). Section 6 discusses the experiments based on synthetic data. Section 7 discusses the case-study using the Walmart sales data. Finally, Section 8 provides discussion about the model of this work along with the directions for future work.

2 RELATED WORK

Our work is closely related to three bodies of work: (censored) demand estimation; matrix completion and estimation; modeling multiple related time series using matrix factorization. We discuss each next.

2.1 (Censored) Demand estimation

Estimating demand is a well-studied problem of interest across several domains. It appears as a sub-problem in the inventory management problems such as the classical news-vendor problem. The distinction between sales and demand data are also well-established in prior works and censoring of demand plays a central role in the most widely studied inventory management problems (e.g. [5], [8], [11]). In [27], the author shows that estimation methods that do not take censoring in to account experience a low-bias problem. In [8], the authors have successfully argued that a lack of visibility (censoring) in the demand data can prove to be costly for inventory planning and that "intelligent analytics" are a valid substitute for the lack of visibility. As such, our work is an instance of "intelligent analytics" to estimate true demand from noisy, censored and missing data.

There are two major approaches to estimating true demand from missing and censored data: Bayesian and data-driven non-parametric. Non-parametric approaches to inferring hidden demand to help with inventory planning have been popular. In a recent work of this flavor non-parametric estimates are determined in an iid setting under censoring [5]. The underlying distribution of interest is assumed to be independent and identically distributed effectively allowing multiple observations of the same distribution and, hence, this is a simpler method than ours. The estimates are shown to be asymptotically optimal in conjunction with a an inventory planning policy. In general, there is a long history of works where the censored demand is estimated in conjunction with a optimal decision-making policy. Works such as [6], [13], [14], [12], [20], [15] solve the inventory management problems by sampling-based policies under censored demand settings. However, these works either consider the iid demand scenarios and then approximate the demand distribution empirically to derive adaptive inventory level decisions for each time step (e.g. [6], [13]), or they use techniques such as stochastic approximations to solve optimization problems for "value" functions that do not rely on *true* demand estimates (e.g. [14], [12], [20]). In [15], the authors use sample average approximations to learn the empirical distributions of demand. In these works, in contrast to our approach, there is little attempt to incorporate other dimensions such as different locations or products, to utilize correlated demand effects which can result in better estimates. Furthermore, stochastic approximations can be unstable and encounter scaling problems [12] which is not the case for us since we use the highly scalable matrix completion and factorization methods.

The Bayesian approach, which is more relevant to our work, assumes a prior probability distribution and computes the MLE estimators of the demand parameters. In [11], the author computes the estimates of the parameter of interest for a Poisson demand distribution and it can be considered an early-precursor to our work. However, only one location (newsstand) across time is considered with the parameter of interest assumed to be identical across time. The author of [19] extends this to the iid Normal case which may not be a good approximation to the reality of sales/demand in the real-world since the demand is non-negative valued and continuously changes. The author of [4] uses the Bayesian approach to estimate unknown parameters with a known prior distribution chosen from the natural conjugate family within an iid setting. Our approach is less restrictive and only assumes the independence across time and locations. In [3], the authors extend the Poisson MLE approach to the setting with substitutes and infer the parameters of interest. However, only a single location is studied. Other works such as [3], [10] and [26] use the Expectation-Maximization approach to infer hidden demand by modeling the demand distribution or customer choice appropriately. In a related approach, the authors of [18] use the multinomial logit model of customer choice for products across different stores and available brands, while the author of [25] assumes that the demand vector for products at time t is a multivariate normal influenced by several observable influencers. This work then uses the EM approach to learn the parameters of interest. Our work assumes nothing about the customer choice and uses other locations (stores) as the second dimension in addition to time. In contrast to all of the above work, our work does not have the limitation of assuming identical distribution across time and allows for distributions to change across time as well as location. Secondly, we use at most one observation per time and location and use no additional product features or customer choice model to garner additional (side) information. Our results naturally extend beyond the Poisson case; we use the Poisson distribution for simplicity and ease of exposition. Lastly, we have provable results about our simple, spectral algorithm unlike the EM algorithm which is excellent procedure but with limited theoretical understanding.

2.2 Matrix completion and estimation.

In a nutshell, the primary conceptual contribution of this work is to identify that the generic censored demand problem is equivalent to the so-called “Latent Variable Model" à la Aldous-Hoover characterization of multi-dimensional exchangeable distributions or what is also known as “Graphons". This connection opens up the possibility of using a minor extension of the result from [7] to devise an estimation algorithm with provable performance guarantees. In lieu of that, our estimation algorithm effectively becomes an instance of “matrix completion and de-noising" based on partial, noisy matrix data. For a detailed discussion on the evolution of the matrix completion, see [7]; for various practical algorithmic implementations, as an example, see [17] and references there in. We note that the significant advantage of using the USVT algorithm of [7] is that it proposes a practical *universal* threshold. Additionally, it does not require symmetric matrices. These are significant advantages over several works in related literature, e.g. [16]. See Section 3.1 for a more detailed discussion.

2.3 Matrix factorization for multiple time series.

A more recent approach, related closely to our work, is to use matrix factorization to de-noise random effects and impute missing information in the censored demand data across a line of products and time. In a recent work [29], the authors factorize the matrix of sales data across products and time. The temporal dependencies are explicitly modeled in an auto-regressive setting. However, censoring and store location based dependencies are not explicitly considered. As such, this matrix-factorization approach is a conceptual extension of online time series prediction with missing data in an auto-regressive setting [2]. Considering the problem as that of multiple (stacked) time series with correlations and dependencies is relevant to our work and considered in previous works such as [9], [21], [22], [28]. As such, [28] is a form of probabilistic matrix factorization (collaborative filtering with latent features) using time as one dimension. Works in probabilistic matrix factorization ([23], [24]) are conceptually close to our work. In [23], for example, Gaussian priors on the matrix are assumed across two dimensions. However, the parameters of the priors are more restrictive than what our approach allows. Our work considers time and locations as the two dimensions of the matrix (for a given product or group of products) but allows each location and time period to have its own independent distributional parameter with no prior knowledge of the parameter value. In that sense, our approach can be regarded as a generalization of these approaches by being able to capture any structure (in the parameters) across the two dimensions of the matrix. Please see Section 8 for a discussion on these generalizations.

3 ALGORITHM

We are given partial observations of the censored demand matrix, $\boldsymbol{X}^p$. 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. $\hat{\boldsymbol{M}}$ of $\boldsymbol{M} = \mathbb{E}[X]$, and (2) extrapolate $\hat{\boldsymbol{M}}$ to obtain $\hat{\Lambda}$ using the knowledge of capacity matrix $\boldsymbol{C}$.

Step 1. Obtaining $\hat{\boldsymbol{M}}$. We apply the Universal Singular Value Thresholding (USVT) of [7] to $\boldsymbol{X}^p$ to obtain $\hat{\boldsymbol{M}}$. For completeness, we describe the USVT algorithm [7]:

(1) Define $b = \max_{i,j} z_{ij}$ and $a = \min_{i,j} z_{ij}$.

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

$$z_{ij} = \begin{cases} X_{ij} & \text{if it is observed in } \boldsymbol{X}^p \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

(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 $\boldsymbol{X}^p$, i.e. empirical estimation of p based on number of entries observed.
(6) Let

$$S = \left\{ i : s_i \geq 2.02\sqrt{T\hat{p}} \right\}. \tag{3}$$

(7) Define

$$W = \frac{1}{\hat{p}} \sum_{i \in S} s_i u_i v_i^T. \tag{4}$$

(8) Let w_{ij} be the (i, j)th element of W. Define

$$\hat{m}_{ij} = \begin{cases} -1 & \text{if } w_{ij} < -1 \\ 1 & \text{if } w_{ij} > 1 \\ w_{ij} & \text{otherwise.} \end{cases} \tag{5}$$

(9) Scale back to the original range:

$$\hat{m}_{ij} \longleftarrow (a+b)/2 + \hat{m}_{ij}(b-a)/2. \tag{6}$$

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

To that end, let us suppose we know $\boldsymbol{M}$ exactly. That is, we know m_{ij} for each $i \leq N, j \leq T$. We also know C_{ij}. Now $m_{ij} = f(\lambda_{ij}, C_{ij})$, where for precise definition of f, please refer to Section 5.1.1. As argued in Lemma 5.1, for any given fixed $C_{ij} \geq 1$, the function f is strictly monotonically increasing in $\lambda_{ij} \in \mathbb{R}_+$. Therefore, 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,

$$\lambda_{ij}^{k+1} = \begin{cases} \frac{1}{2}\left(\lambda_{ij}^{UB} + \lambda_{ij}^{LB}\right), & \text{if } \lambda_{ij}^{UB} < \infty \\ 2\lambda_{ij}^{LB}, & \text{if } \lambda_{ij}^{UB} = \infty. \end{cases} \tag{7}$$

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

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 $\boldsymbol{\Lambda}$ by $\hat{\boldsymbol{\Lambda}}$.

3.1 Universal Thresholding

We note that *Step 1* of the algorithm could be replaced by other competing singular value thresholding algorithms and heuristics. However, there are significant advantages of using the USVT algorithm of [7]: the choice of the threshold is *universal* which is in contrast to many algorithms in literature which do not specify a principled approach to choosing the threshold. Secondly, the USVT algorithm allows us to establish attractive asymptotic properties of the $MSE(\hat{\boldsymbol{M}})$ and $MSE(\hat{\boldsymbol{\Lambda}})$ as shown in Section 4. As a reference, we compare the USVT algorithm to a similar spectral algorithm described in [16] which is applicable to latent variable models for the generalized stochastic block models (GSBM). We first note that the algorithm in [16] is applicable to symmetric matrices while our setting does not have such a restriction. Additionally, the algorithm in [16] does not specify a way to choose the threshold. We use the Appendix B to show a comparison of the $MSE(\hat{\boldsymbol{M}})$ for symmetric matrices using the USVT algorithm and the algorithm from [16] (with the same number of eigenvalues retained as those in the USVT algorithm). We note that both algorithms have similar performance. However, the USVT algorithm is always the better option.

4 MAIN RESULT

4.1 Operating assumptions

We note the key model assumptions before stating the main result. Let Y_{ij} be true demand at store $i \in [N]$ at time $j \in [T]$. Y_{ij} is an independent random variable with Poisson distribution whose mean is λ_{ij}. Each store $i \in [N]$ has latent feature $\theta_i \in \Omega_1$ associated with it. Each time $j \in [T]$ has latent feature $\rho_j \in \Omega_2$ associated with it. We shall assume that Ω_1 and Ω_2 are compact sets in finite dimensional Euclidian space. For concreteness and simplicity, let us suppose $\Omega_1 = \Omega_2 = [0, 1]^d$ for some finite $d \geq 1$. We assume that $\lambda_{ij} = h(\theta_i, \rho_j)$, where $h : [0, 1]^d \times [0, 1]^d \longrightarrow \mathbb{R}_+$ is a Lipschitz function with Lipschitz constant L. Given these assumptions, it immediately follows that there exists $\lambda^* \in \mathbb{R}_+$ so that $\sup_{\theta, \rho \in [0,1]^d} h(\theta, \rho) = \lambda^*$. We note that our Lipschitz assumption imposes a more realistic structure than the standard low-rank assumption in literature. We discuss the specific implications of this assumption in Section 4.3.

We assume that the inventory capacity, C_{ij} at store $i \in [N]$ and time $j \in [T]$ is a random variable whose distribution is parametrized by θ_i and ρ_j. Specifically, $\mathbb{P}(C_{ij} = k) = g_k(\theta_i, \rho_j)$ with $g_k : [0, 1]^d \times [0, 1]^d \longrightarrow [0, 1]$ is a Lipschitz function with Lipschitz constant L_k. We assume that maximum capacity is bounded above by a universal constant C^*, i.e. $C_{ij} \leq C^*$ with probability 1 for all $i \in [N], j \in [T]$. With that said, we assume that all realized capacity values are known. This is a realistic assumption because most modern retailers have equipped themselves with the ability to record and access precise inventory information. The censored demand realized at store i at time j is $X_{ij} = \min(Y_{ij}, C_{ij})$. Let $m_{ij} = \mathbb{E}[X_{ij}]$. Each X_{ij} is observed with probability $p \in (0, 1]$, independently.

4.2 Statement of main result

The main result is about the performance of the algorithm described in Section 3 in terms of its ability to estimate $\hat{\boldsymbol{\Lambda}}$. As stated, the algorithm has two estimation steps. Therefore, we state results about the estimation error introduced in each step. Stitching them together will lead to the main result.

Estimation Error in $\hat{\boldsymbol{M}}$*.* We state a bound on $MSE(\hat{\boldsymbol{M}})$ induced by the Step 1 (USVT) of the algorithm.

Lemma 4.1. *For a given $\varepsilon \in (0,1)$, let $p \geq N^{-1+\varepsilon}$. When N is large enough for a given ε, we have*

$$MSE(\hat{\boldsymbol{M}}) \leq c_1 \frac{N^{-\frac{1}{d+2}}}{\sqrt{p}}, \tag{8}$$

where $c_1 = \alpha C^* \big(1 + C^* \Gamma(L_\psi, d)\big)$ and $\Gamma(d, L_\psi) = (4dL_\psi)^{d/2}$, and where L_ψ is a Lipscthiz constant and d is the dimension of the latent variable space. α is a universal constant.

Estimation Error in $\hat{\boldsymbol{\Lambda}}$: using $\hat{\boldsymbol{M}}$. We state a bound on the $MSE(\hat{\boldsymbol{\Lambda}})$ induced by Step 2 of the algorithm.

Lemma 4.2. *For any $i \in [N]$, $j \in [T]$,*

$$|\hat{\lambda}_{ij} - \lambda_{ij}| \leq \frac{|\hat{m}_{ij} - m_{ij}|}{\mathbb{P}(Q \leq \max(0, C_{ij} - 2))}, \tag{9}$$

where Q is Poisson random variable with parameter $\tilde{\lambda}_{ij} = \max(\lambda_{ij}, \hat{\lambda}_{ij})$.

For any $\tilde{\lambda}_{ij}$, $\mathbb{P}(Q \leq \max(0, C_{ij} - 2)) \geq \mathbb{P}(Q = 0) = \exp(-\tilde{\lambda}_{ij})$. Since $\max_{ij} \lambda_{ij} \leq \lambda^*$, it follows that

$$|\hat{\lambda}_{ij} - \lambda_{ij}| \leq \exp(\lambda^*)|\hat{m}_{ij} - m_{ij}| \tag{10}$$

That is,

$$MSE(\hat{\boldsymbol{\Lambda}}) \leq \exp(2\lambda^*) MSE(\hat{\boldsymbol{M}}). \tag{11}$$

Putting It Together. From Theorems 4.1 and 4.2, we obtain the following result.

Theorem 4.3. *For a given $\varepsilon \in (0,1)$, let $p \geq N^{-1+\varepsilon}$. When N is large enough, for a given ε, we have*

$$MSE(\hat{\boldsymbol{\Lambda}}) \leq c_1 \exp\{2\lambda^*\} \frac{N^{-\frac{1}{d+2}}}{\sqrt{p}}, \tag{12}$$

where $c_1 = \alpha C^* \big(1 + C^* \Gamma(L_\psi, d)\big)$ and $\Gamma(d, L_\psi) = (4dL_\psi)^{d/2}$, and where L_ψ is a Lipscthiz constant and d is the dimension of the latent variable space. λ^* is as defined in Lemma 4.2 above. α is a universal constant.

As a consequence, as long as $p \gg N^{\frac{-2}{d+2}}$ we have $MSE(\hat{\boldsymbol{\Lambda}}) \longrightarrow 0$ as $N \longrightarrow \infty$.

4.3 Implications

Theorem 4.3 captures the fact that with enough samples and well-behaved constants, as $N \longrightarrow \infty$, the errors in both steps of the algorithm go to 0. It is the error in Step 2 that should be affected by censoring. What is surprising is that even when $C_{ij} = 1$ for all $i \in [N], j \in [T]$, in the regime mentioned above, error goes to 0! That is, if effectively there is *only one* product on the shelf, knowing whether it is purchased or not is sufficient to estimate the entire demand rate!

As we pay closer attention to Lemma 4.2, notice that as $C_{ij} \longrightarrow \infty$, the error in $\hat{\lambda}_{ij}$ converges to error in $\hat{m}_{ij}$. In other words, as censoring reduces, the censoring induced error in the Step 2 of the algorithm reduces – naturally, as one would expect. And vice versa. This expected qualitative behavior gives us confidence in the fact that the bounds on the estimation error are capturing first-order effects.

We now compare Lemma 4.1 to the standard results in literature which assume that the latent mean matrix, $\boldsymbol{M}$, is low-rank. First, note that the Lipschitz assumption is a strict generalization of the low-rank assumption because the latter can be viewed as a "specific" Lipschitz function. Consequently, this allows greater flexibility in capturing "reality" using the Lipschitz structure. However, greater model flexibility comes at a cost. This cost is the greater amounts of data required for estimation when using the more general Lipschitz setting. Specifically, in the discussion proceeding Theorem

2.1 in [7], it is shown that for consistent estimation of $\boldsymbol{M}$, as $N \longrightarrow \infty$, it is necessary for $p \gg \frac{r}{N}$. In contrast, the Lipschitz structure comes at the cost of needing to observe more data, for a fixed d: in our setting, asymptotic consistency is achieved if $p \gg N^{-2/(d+2)}$.

5 PROVING THE RESULT

5.1 Preliminaries

Here we establish a few useful preliminary properties that will be utilized in establishing the proof of our main result. We first determine the mean of a truncated Poisson distribution and using some helpful properties then establish a relationship between the means of the truncated and corresponding non-truncated Poisson distributions. Next, we establish that the mean matrix, $\boldsymbol{M}$, is Lipschitz in its latent parameters which allows us to bound the nuclear norm, $||\boldsymbol{M}||_*$. The relationship between the means of the truncated and non-truncated Poisson distributions and the bound on the nuclear norm of the matrix of means of the truncated Poisson random variables, $||\boldsymbol{M}||_*$, will then allow us to establish our main result in Sections 5.2 and 5.3.

5.1.1 Mean of a truncated (censored) Poisson random variable. Consider a Poisson random variable, say Q such that $\mathbb{E}[Q] = \lambda$. For any $C \geq 1$, let the truncation of Q at C be denoted as R, that is,

$$R = \min(Q, C). \tag{13}$$

Let

$$\begin{aligned} m &\equiv \mathbb{E}[R] \\ &= \sum_{t=0}^{C-1} t\mathbb{P}(R = t) + C\mathbb{P}(R = C) \\ &= \sum_{t=0}^{C-1} t\mathbb{P}(Q = t) + C\big(\sum_{t=C}^{\infty} \mathbb{P}(Q = t)\big) \\ &= \sum_{t=0}^{\infty} t\mathbb{P}(Q = t) - \sum_{t=C}^{\infty} (t - C)\mathbb{P}(Q = t) \\ &= \mathbb{E}[Q] - \sum_{t=C}^{\infty} (t - C)\mathbb{P}(Q = t) \\ &= \lambda - \sum_{t=C}^{\infty} (t - C)\frac{\exp(-\lambda)\lambda^t}{t!} \\ &\equiv f(\lambda, C). \end{aligned} \tag{14}$$

That is, $m = f(\lambda, C)$. This function f satisfies the following useful properties.

Lemma 5.1. *The non-negative valued function $f : \mathbb{R}_+ \times \mathbb{Z}_+ \longrightarrow \mathbb{R}_+$, as defined in (14), satisfies the following: for any $\lambda \in \mathbb{R}_+$ and $C \in \mathbb{Z}_+$,*

$$\frac{\partial f}{\partial \lambda}(\lambda, C) = \mathbb{P}(Q \leq \max(0, C - 2)) \leq 1. \tag{15}$$

(15) appeals to our intuition where the derivative with respect to λ is small and positive in situations where there is a high degree of censoring (small C). In such situations, the truncated mean, m, will increase very slowly as λ is increased. On the other hand, in situations where there is little to no censoring, i.e. C is large, we expect the truncated mean to approximate the un-truncated mean,

λ, which will grow much more rapidly as λ increases. Note that the derivative remains positive, bounded above by 1 and bounded below by $\exp(-\lambda)$, under all circumstances.

Proof. To start with, consider case when $C = 1$. Then,

$$f(\lambda, 1) = 1 - \exp(-\lambda). \tag{16}$$

In this case,

$$\frac{\partial f}{\partial \lambda}(\lambda, 1) = \exp(-\lambda) = \mathbb{P}(Q \leq 0). \tag{17}$$

Therefore, when $C = 1$, for any $\lambda \in \mathbb{R}_+$, we have

$$\frac{\partial f}{\partial \lambda}(\lambda, 1) = \mathbb{P}(Q \leq \max(0, C - 2)), \tag{18}$$

where Q is Poisson random variable with parameter λ.
Now we consider scenario where $C \geq 2$. We start by deriving the precise form of $\frac{\partial f}{\partial \lambda}(\lambda, C)$. To that end, we shall use the following definition:

$$f(\lambda, C) = \sum_{t=0}^{C-1} t \exp(-\lambda) \frac{\lambda^t}{t!} + \sum_{t=C}^{\infty} C \exp(-\lambda) \frac{\lambda^t}{t!}. \tag{19}$$

Therefore,

$$\begin{aligned}\frac{\partial f}{\partial \lambda}(\lambda, C) = &\sum_{t=0}^{C-1} \frac{t \exp(-\lambda)}{t!} \Big(t\lambda^{t-1} - \lambda^t \Big) \\ &+ C \sum_{t=C}^{\infty} \frac{\exp(-\lambda)}{t!} \Big(t\lambda^{t-1} - \lambda^t \Big).\end{aligned} \tag{20}$$

Consider the first term in (20):

$$\begin{aligned}&\sum_{t=0}^{C-1} \frac{t \exp(-\lambda)}{t!} \Big(t\lambda^{t-1} - \lambda^t \Big) \\ &= \exp(-\lambda) \Big(\sum_{t=1}^{C-1} \frac{t\lambda^{t-1}}{(t-1)!} - \sum_{t=1}^{C-1} \frac{\lambda^t}{(t-1)!} \Big) \\ &= \exp(-\lambda) \Big(\sum_{t=1}^{C-1} \frac{(t-1)\lambda^{t-1}}{(t-1)!} + \sum_{t=1}^{C-1} \frac{\lambda^{t-1}}{(t-1)!} - \sum_{t=1}^{C-1} \frac{\lambda^t}{(t-1)!} \Big) \\ &= \exp(-\lambda) \Big(\sum_{t=1}^{C-2} \frac{\lambda^t}{(t-1)!} + \sum_{t=1}^{C-1} \frac{\lambda^{t-1}}{(t-1)!} - \sum_{t=1}^{C-1} \frac{\lambda^t}{(t-1)!} \Big) \\ &= \Big(\sum_{t=0}^{C-2} \exp(-\lambda) \frac{\lambda^t}{t!} \Big) - \Big(\exp(-\lambda) \frac{\lambda^{C-1}}{(C-2)!} \Big) \\ &= \mathbb{P}(Q \leq C-2) - \exp(-\lambda) \frac{\lambda^{C-1}}{(C-2)!},\end{aligned} \tag{21}$$

where Q is Poisson random variable with mean λ.

Consider the second term in (20):

$$
\begin{aligned}
C\sum_{t=C}^{\infty}\frac{\exp(-\lambda)}{t!}\left(t\lambda^{t-1}-\lambda^{t}\right)& \\
&= C\exp(-\lambda)\left(\sum_{t=C}^{\infty}\frac{\lambda^{t-1}}{(t-1)!}-\frac{\lambda^{t}}{t!}\right)\\
&= \exp(-\lambda)\frac{\lambda^{C-1}}{(C-2)!}. \qquad (22)
\end{aligned}
$$

Using (21) and (22) in (20), we obtain

$$\frac{\partial f}{\partial\lambda}(\lambda,C)=\mathbb{P}(Q\le C-2). \qquad (23)$$

From (18) and (23), we have that for all $\lambda\in\mathbb{R}_+$ and $C\in\mathbb{Z}_+$,

$$\frac{\partial f}{\partial\lambda}(\lambda,C)=\mathbb{P}(Q\le\max(0,C-2)), \qquad (24)$$

where Q is a Poisson random variable with parameter λ. This completes the proof of Lemma. □

5.1.2 Sensitivity analysis of $\mathbf{\Lambda}$ *with respect to* $\boldsymbol{M}$. We state the following result regarding sensitivity analysis of f^{-1} as defined in (14).

Lemma 5.2. *Given fixed $C\in\mathbb{Z}$, let (m_1,λ_1) and (m_2,λ_2) be pairs of means of truncated Poisson and Poisson random variables. That is, $m_k=f(\lambda_k,C)$ for $k=1,2$ with f as defined in (14). Then,*

$$|\lambda_1-\lambda_2|\le\frac{|m_1-m_2|}{\mathbb{P}(Q\le\max(0,C-2))}, \qquad (25)$$

where Q is Poisson random variable with parameter $\lambda=\max(\lambda_1,\lambda_2)$.

Proof. Without loss of generality, let us assume that $m_1\le m_2$ and hence $\lambda_1\le\lambda_2$. Given fixed $C\in\mathbb{Z}$, the function f maps $\lambda\in\mathbb{R}_+$ to $m\in\mathbb{R}_+$. Let g be the inverse of the map, i.e. inverse of $f(\lambda,C)$ with respect to first argument keeping second argument fixed. Therefore, $g(m_k)=\lambda_k$ for $k=1,2$. We know that f is continuous, differentiable and strictly monotonic over $\mathbb{R}_+$. Therefore, g is continuous and differentiable as well. Then

$$
\begin{aligned}
|\lambda_1-\lambda_2| &= |g(m_1)-g(m_2)|\\
&= |g'(m)||m_1-m_2|, \qquad (26)
\end{aligned}
$$

where the above equality follows from the Mean-Value Theorem with $g'(\cdot)$ being the derivative of g, and $m\in(m_1,m_2)$. Since f and g both are differentiable over $\mathbb{R}_+$, by elementary argument in analysis, it follows that

$$|g'(m)|=\frac{1}{|f'(\lambda)|} \qquad (27)$$

where λ is such that $f(\lambda,C)=m$ and $f'(\lambda)=\frac{\partial f}{\partial\lambda}(\lambda,C)$. Due to monotonicity of f, it follows that $\lambda\in(\lambda_1,\lambda_2)$. Substituting (27) in (26), and using Lemma 5.1, we obtain

$$|\lambda_1-\lambda_2|=\frac{|m_1-m_2|}{\mathbb{P}(Q\le\max(0,C-2))}, \qquad (28)$$

where Q is Poisson random variable with parameter $\lambda \in (\lambda_1, \lambda_2)$. It can be easily verified that $\mathbb{P}(Q \leq \max(0, C-2))$ is a monotonically decreasing function of λ for a fixed C. Therefore, for all $\lambda \in (\lambda_1, \lambda_2)$, it is bounded below by $\lambda = \lambda_2$. Therefore, we conclude that

$$|\lambda_1 - \lambda_2| \leq \frac{|m_1 - m_2|}{\mathbb{P}(Q \leq \max(0, C-2))}, \tag{29}$$

where Q is Poisson random variable with parameter $\lambda = \max(\lambda_1, \lambda_2)$. This completes the proof of Lemma 5.2. □

5.1.3 Lipschitz property of $\boldsymbol{M}$. Next we establish that, $m_{ij} = \mathbb{E}[X_{ij}]$, the (i, j)th element of $\boldsymbol{M}$, is a Lipschitz function of the latent features θ_i and ρ_j associated with store $i \in [N]$ and time $j \in [T]$.

Lemma 5.3. *Let the assumptions stated in Section 4.1 hold. Then, there exists a Lipschitz function* $\psi : [0,1]^d \times [0,1]^d \longrightarrow [0, C^*]$ *so that* $m_{ij} = \psi(\theta_i, \rho_j)$ *for* $i \in [N],\ j \in [T]$.

Proof. By definition,

$$\begin{aligned} m_{ij} = \mathbb{E}[X_{ij}] &= \mathbb{E}[\mathbb{E}[X_{ij}|C_{ij}]] \\ &= \sum_{k=1}^{C^*} \mathbb{E}[X_{ij}|C_{ij} = k]\mathbb{P}(C_{ij} = k). \end{aligned} \tag{30}$$

Now given $C_{ij} = k$, $\mathbb{E}[X_{ij}|C_{ij} = k]$ is precisely $f(\lambda_{ij}, k)$ where f is defined in (14). By the assumptions of Section 4.1, $\lambda_{ij} = h(\theta_i, \rho_j)$ and $\mathbb{P}(C_{ij} = k) = g_k(\theta_i, \rho_j)$. Therefore,

$$m_{ij} = \sum_{k=1}^{C^*} f(h(\theta_i, \rho_j), k)\, g_k(\theta_i, \rho_j) \qquad \equiv \psi(\theta_i, \rho_j). \tag{31}$$

Next, we establish that ψ is a Lipschitz function. To that end, by Lemma 5.1, $f(\cdot, k)$ is a Lipschitz function with Lipschitz constant 1 in it's first argument for all $k \geq 1$. By assumption of Section 4.1, h is a Lipschitz function with constant L. Therefore, for a fixed k, $f(h(\theta_i, \rho_j), k)$ is a Lipschitz function of (θ_i, ρ_j) with Lipschitz constant L.
By the assumptions of Section 4.1, g_k is a Lipschitz function with constant L_k for $1 \leq k \leq C^*$.
The following are easy to verify compositional rules associated with Lipschitz functions:

(1) If ϕ_1 and ϕ_2 are Lipschitz functions with constants z_1 and z_2, respectively, then $\phi_3 = \phi_1 + \phi_2$ is a Lipschitz function with Lipschitz constant $z_3 = z_1 + z_2$.
(2) ϕ_1 and ϕ_2 are Lipschitz functions with constants z_1 and z_2, then $\phi_3 = \phi_1 \times \phi_2$ is also a Lipschitz function with Lipschitz constant $z_1|\phi_2|_\infty + z_2|\phi_1|_\infty$.

Note that $|f(\cdot, k)| \leq k$ and $k \leq C^*$, that is, $|f(\cdot, k)|_\infty \leq C^*$; and by definition $|g_k|_\infty \leq 1$. Therefore, by putting all of the above discussion together, we obtain that ψ is a Lipschitz continuous function with Lipschitz constant L_ψ, where

$$L_\psi \leq C^* \Big(L + \sum_{k=1}^{C^*} L_k \Big). \tag{32}$$

This completes the proof of Lemma 5.3. □

5.1.4 Bounding $\|\boldsymbol{M}\|_*$. We shall utilize the Lipschitz property of $\boldsymbol{M}$ established in Lemma 5.3 to bound the nuclear norm of $\|\boldsymbol{M}\|_*$ as stated in Lemma 5.4 below. The proof of the Lemma below is a straightforward adaption of the arguments from [7, Lemma 3.6]. We present them here for completeness.

Lemma 5.4. *Given $\boldsymbol{M}$ as defined above, for any small enough $\delta > 0$,*

$$\|\boldsymbol{M}\|_* \leq \delta N\sqrt{T} + C^*\Gamma(L_\psi, d)\sqrt{NT\delta^{-d}}, \tag{33}$$

where $\Gamma(L_\psi, d)$ is a constant that depends on Lipschitz constant L_ψ of ψ as defined in (32) and d is the dimension of the latent space. The constant $C^* = \|\psi\|_\infty$.

Proof. By Lemma 5.3, the (i, j)th element of $\boldsymbol{M}$, $m_{ij} = \psi(\theta_i, \rho_j)$ where ψ is Lipschitz in its arguments and $\psi : [0, 1]^d \times [0, 1]^d \to [0, C^*]$. For any given $\delta > 0$, it is easy to see that one can find a finite covering $P_1(\delta)$ and $P_2(\delta)$ of $[0, 1]^d$ so that for any $\theta, \rho \in [0, 1]^d$, there exists $\theta' \in P_1(\delta)$ and $\rho' \in P_2(\delta)$ so that

$$|\psi(\theta, \rho) - \psi(\theta', \rho')| \leq \delta. \tag{34}$$

For example, let $\zeta = \lceil \frac{2dL_\psi}{\delta} \rceil$ and define $P_1(\delta) = P_2(\delta) = P(\delta)$, where

$$P(\delta) = \{(k_1/\zeta, \ldots, k_d/\zeta) : k_1, \ldots, k_d \in [\zeta]\}. \tag{35}$$

Then, for any $\theta, \rho \in [0, 1]^d$, we can find $\theta', \rho' \in P(\delta)$ so that

$$\|(\theta, \rho) - (\theta', \rho')\|_2 \leq \|(\theta, \rho) - (\theta', \rho')\|_1 \leq \frac{2d}{\zeta} \leq \frac{\delta}{L_\psi}. \tag{36}$$

Therefore, by Lipschitz property of ψ, we have that $|\psi(\theta, \rho) - \psi(\theta', \rho')| \leq \delta$ as desired. In this construction, we have

$$|P(\delta)| \sim \zeta^d \leq \Gamma_1(d, L_\psi)\delta^{-d}, \tag{37}$$

where $\Gamma_1(d, L_\psi) = (4dL_\psi)^d$.
For latent feature θ_i corresponding to store $i \in [N]$, find closest element in $P(\delta)$, and let it denote by $p_1(\theta_i)$. Similarity, for latent feature ρ_j corresponding to time $j \in [T]$, find closest element in $P(\delta)$, and let it denote by $p_2(\rho_j)$. Create matrix $B = [b_{ij}]$ where $b_{ij} = \psi(p_1(\theta_i), p_2(\rho_j))$. As argued above, we have that for all $i \in [N], j \in [T]$

$$|m_{ij} - b_{ij}| \leq \delta. \tag{38}$$

Therefore,

$$\|\boldsymbol{M} - B\|_F \leq \delta\sqrt{NT}. \tag{39}$$

This gives us

$$\begin{aligned}\|\boldsymbol{M}\|_* &\leq \|\boldsymbol{M} - B\|_* + \|B\|_* \\ &\leq \sqrt{N}\|\boldsymbol{M} - B\|_F + \|B\|_* \\ &\leq \delta N\sqrt{T} + \|B\|_*. \end{aligned} \tag{40}$$

In above, we used the inequality that for any real-valued matrix Q, $\|Q\|_* \leq \sqrt{\text{rank}(Q)}\|Q\|_F$. We shall use the same inequality again to bound $\|B\|_*$. To obtain a tight bound, let us argue that the rank of B does not scale with N and T. To that end, consider any two columns, say $j, j' \in [T]$. If $p_2(\rho_j) = p_2(\rho_{j'})$, then it follows that the columns j and j' of B are identical. That is, there are can be at most $|P(\delta)|$ distinct columns of B. Similarly, there can be at most $|P(\delta)|$ distinct rows of B. That is, $\text{rank}(B) \leq |P(\delta)|$. Finally, we know that $\|\psi\|_\infty \leq C^*$. Therefore, we have

$$\begin{aligned}\|B\|_* &\leq \sqrt{|P(\delta)|}\|B\|_F \leq \sqrt{|P(\delta)|}\sqrt{NT}\|\psi\|_\infty \\ &\leq \sqrt{|P(\delta)|}\sqrt{NT}C^*. \end{aligned} \tag{41}$$

Putting everything together, we have

$$\|\boldsymbol{M}\|_* \leq \delta N\sqrt{T} + C^*\sqrt{\Gamma_1(L_\psi, d)}\sqrt{NT\delta^{-d}}. \tag{42}$$

where $\Gamma_1(d, L_\psi) = (4dL_\psi)^d$, as in (37). □

An immediate implication of the above Lemma is that by selecting $\delta = N^{-\frac{1}{d+2}}$, we obtain

$$\|\boldsymbol{M}\|_* \leq \big(1 + C^*\Gamma(L_\psi, d)\big)\sqrt{T}N^{1-\frac{1}{d+2}}. \tag{43}$$

where $\Gamma(d, L_\psi) = (4dL_\psi)^{d/2}$.

5.2 Key Enabler

We state the key enabler [7, Theorem 2.1]. We state it here for non-normalized setup as described below and applicable to our setting. Consider an $m \times n$ matrix $A = [A_{ij}]$ of interest. Let $A_{ij} \in [-B, B]$ for all $i \in [m], j \in [n]$ for some $B \geq 1$. Let $m \leq n$. Let $Z = [Z_{ij}]$ be an $m \times n$ random matrix whose entries are independent such that $\mathbb{E}[Z_{ij}] = A_{ij}$ and $Z_{ij} \in [-B, B]$ with probability 1. Each entry of the matrix Z is observed independently with probability $p \in [0, 1]$ and unobserved with probability $1 - p$. The Universal Singular Value Thresholding (USVT) algorithm as described in Section 3 when applied to Z produces an estimation matrix $\hat{A}$. The expected mean squared error is defined as

$$\mathsf{MSE}(\hat{A}) = \frac{1}{mn}\mathbb{E}\big[\|\hat{A} - A\|_F^2\big]. \tag{44}$$

Then, as claimed and proved in [7],

Theorem 5.5 (Theorem 2.1 of [7]). *Let there be a given $\varepsilon > 0$. Suppose $p \geq n^{-1+\varepsilon}$. Then*

$$\mathsf{MSE}(\hat{A}) \leq \alpha \min\left\{B\frac{\|A\|_*}{m\sqrt{np}} + \frac{B^2}{np}, \frac{\|A\|_*^2}{mn}, B^2\right\} + B^2\beta(\varepsilon)\exp(-\gamma np), \tag{45}$$

where α and γ are universal constants and $\beta(\varepsilon)$ depends on ε.

5.3 Proof of Lemmas 4.1 and 4.2, Theorem 4.3

Proof of Lemma 4.1. The application of Theorem 5.5 (where A is M, $B = C^*$, $m = N$ and $n = T$), we find that as long as $p \geq N^{-1+\varepsilon} \geq T^{-1+\varepsilon}$ for any $0 < \varepsilon < 1$, for N large enough, the Step 1 of our algorithm described in Section 3 produces $\hat{\boldsymbol{M}}$ so that

$$\mathsf{MSE}(\hat{\boldsymbol{M}}) \leq \alpha\left(\frac{C^*\|\boldsymbol{M}\|_*}{N\sqrt{Tp}} + \frac{(C^*)^2}{Tp}\right) + (C^*)^2\beta(\varepsilon)\exp(-\gamma Tp), \tag{46}$$

By plugging in bound from (43) and using $T \geq N$, we obtain

$$\mathsf{MSE}(\hat{\boldsymbol{M}}) \leq c_1\frac{N^{-\frac{1}{d+2}}}{\sqrt{p}} + \frac{c_2}{Tp} + c_3\exp(-N^\varepsilon). \tag{47}$$

where $c_1 = \alpha C^*\big(1 + C^*\Gamma(L_\psi, d)\big)$, $c_2 = \alpha(C^*)^2$ and c_3 depends on ε and γ. As earlier, we have that $\Gamma(d, L_\psi) = (4dL_\psi)^{d/2}$, where d is the dimension of the latent space and L_ψ is a Lipschitz constant. Since $p \geq N^{-1+\varepsilon}$, as N scales, the first term on the right is dominant, leading to

$$\mathsf{MSE}(\hat{\boldsymbol{M}}) \leq c_1\frac{N^{-\frac{1}{d+2}}}{\sqrt{p}}, \tag{48}$$

Proof of Lemma 4.2. Lemma 4.2 follows immediately from Lemma 5.2.

Proof of Theorem 4.3. The proof of Theorem 4.3 follows immediately by putting together Lemma 4.1 and implication (11) of Lemma 4.2.

6 SIMULATED EXPERIMENTS

6.1 Experimental Setup

We conduct simulated experiments to establish the various properties of the estimates $\hat{\boldsymbol{M}}$ and $\hat{\boldsymbol{\Lambda}}$. We consider the following metrics of evaluation: $RMSE(\hat{\boldsymbol{M}})$, $RMSE(\hat{\boldsymbol{\Lambda}})$ and the Ratio: $\frac{RMSE(\hat{\boldsymbol{\Lambda}})}{RMSE(\hat{\boldsymbol{M}})}$. This last quantity, Ratio, helps establish the relationship between $RMSE(\hat{\boldsymbol{\Lambda}})$ and $RMSE(\hat{\boldsymbol{M}})$ to confirm the various implications of Lemma 4.2 and Theorem 4.3, as discussed in Section 4.3.

For our experiments, θ_i and ρ_j are randomly sampled from a $\mathbb{U}(1)$ uniform distribution for all $1 \leq i \leq N, 1 \leq j \leq T$, unless noted otherwise. The (hidden, unknown) parameters of interest, λ_{ij} are determined using the following Lipschitz function: $\lambda_{ij} = \boldsymbol{h}(\theta_i, \rho_j) = \frac{J}{\exp\{-\theta_i - \rho_j - \alpha\theta_i\rho_j\}}$. Note that $\alpha \in (0, 1)$ is random but fixed constant. J is a scaling constant used to generate as large values as needed for a simulation. Therefore, in all the experiments discussed, comparisons are only drawn for a constant scaling constant, unless explicitly stated otherwise. The stocking level realizations of C_{ij} are known for all $1 \leq i \leq N, 1 \leq j \leq T$. The random, but unknown, matrix of *true* demand values is sampled as $Y_{ij} \sim Poisson(\lambda_{ij})$ for all $1 \leq i \leq N, 1 \leq j \leq T$. Each demand realization is then subject to censoring due to the stocking level C_{ij}. This gives us the matrix $\boldsymbol{X}$. We fix the probability of observation, p. Using that, we observe each entry of the matrix X independently with probability p giving us the matrix X^p. We then estimate the censored means and (hidden) original parameters using the algorithm described in Section 3.

The simulation experiments are designed to help explore various properties of the results stated in Section 4. We would like the experiments to reveal how our evaluation metrics are affected by the amounts of censoring. As discussed in Section 4.3, we expect the Ratio to decay to a value of 1 as the degree of censoring reduces, and be increasingly greater than 1 as censoring increases. Additionally, we expect that as the probability of observation, p, is increased the estimates improve. We also expect to confirm the consistency property of both $RMSE(\hat{\boldsymbol{M}})$ and $RMSE(\hat{\boldsymbol{\Lambda}})$. Finally, we would also like to study the impact of *structure* on the estimates. We intuitively expect that the more *structure* there is to exploit, the better the estimates will be.

6.2 Effects of Censoring and Probability of Observation

This set of simulated experiments show the effect of censoring on the Ratio: $\frac{RMSE(\hat{\boldsymbol{\Lambda}})}{RMSE(\hat{\boldsymbol{M}})}$ and RMSE($\hat{\boldsymbol{M}}$) and how they vary across different levels of p. The parameter scaling constant, $J = 15$. To illustrate the effects of censoring clearly, all C_{ij} are kept the same for each experiment and denoted by C. For this set of experiments we used a matrix size of 10,000.

Figure 1 shows that as the censoring levels decrease, i.e. C increases, the Ratio decreases and plateaus out to equal 1 for all values of p used. At higher levels of censoring, i.e. C is smaller, the Ratio is larger, as expected. This behavior holds across all values of p.

Figure 2, shows that different levels of p result in quantitatively different profiles of $RMSE(\hat{\boldsymbol{M}})$. The higher the value of p, the lower the $RMSE(\hat{\boldsymbol{M}})$), as we would expect. Also we note that $RMSE(\hat{\boldsymbol{M}})$ increases as censoring decreases. This makes sense because m_{ij} are the censored means and censoring reduces the range of possible values (support) of m_{ij}. Therefore, it makes sense that the $RMSE(\hat{\boldsymbol{M}})$ is smaller in situations of increased censoring.

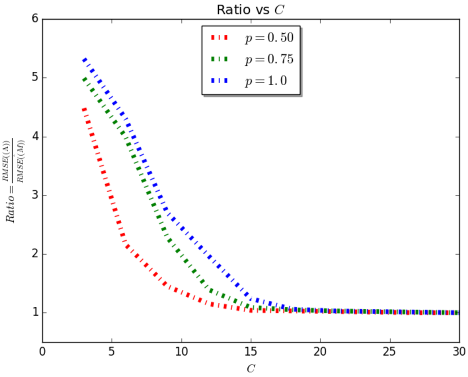

Fig. 1. The effect of decreasing censoring (varying C) on the Ratio for different levels of p. $J = 15$.

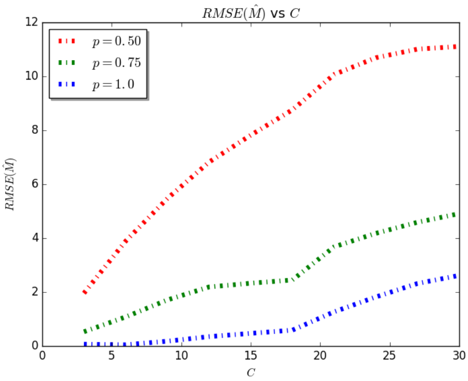

Fig. 2. The effect of descreasing censoring (varying C) on the RMSE($\hat{\boldsymbol{M}}$) for different levels of p. $J = 15$.

6.3 Consistency of Estimates: $\hat{M}$ and $\hat{\Lambda}$

For this series of simulations, we used the scaling constant $J = 5$. In order to study the effect of the levels of censoring we vary it across experiments but keep all C_{ij} constant within each experiment. We explore three levels of censoring: **significant** ($C = 2$), **mild** ($C = 5$) and **little**($C = 10$).

Figures 3 and 4 show the results of this set of experiments. As expected, we note that $RMSE(\hat{\boldsymbol{M}})$ and $RMSE(\hat{\boldsymbol{\Lambda}})$ both decrease as the size of the matrix increases. As claimed in Section 4.3, $RMSE(\boldsymbol{\Lambda})$ is consistently smaller when the levels of censoring are smaller (higher C values). Note that the effect is exactly the opposite for $RMSE(\boldsymbol{M})$, as argued in Section 6.2. However, as the matrix size increases, $RMSE(\boldsymbol{M})$ decreases for all levels of censoring.

Figure 5 shows that the Ratio is lowest when there is little censoring (≈ 1) and increases in the presence of increased censoring. However, the size of the matrix appears to have minimal impact on the Ratio for a fixed level of censoring. This is an appealing property of the Ratio because both $RMSE(\hat{\boldsymbol{M}})$ and $RMSE(\hat{\boldsymbol{\Lambda}})$ do get affected by the size of the matrix. However, their ratio does not indicating that the changes in $RMSE(\hat{\boldsymbol{M}})$ and $RMSE(\hat{\boldsymbol{\Lambda}})$ are correlated.

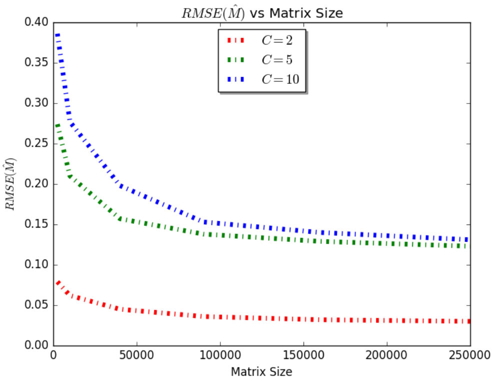

Fig. 3. Effect of increasing the size of the matrix on the RMSE($\hat{\boldsymbol{M}}$). The different plots represent different levels of censoring. $J = 5$. The three levels of censoring are: little ($C = 10$), mild ($C = 5$) and significant ($C = 2$).

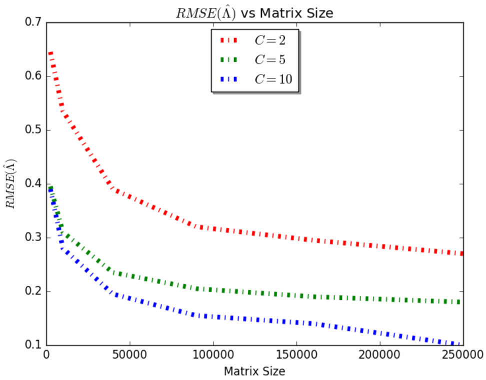

Fig. 4. Effect of increasing the size of the matrix, X^p, on the RMSE($\hat{\Lambda}$). The different plots represent different levels of censoring. $J = 5$. The three levels of censoring are: little ($C = 10$), mild ($C = 5$) and significant ($C = 2$).

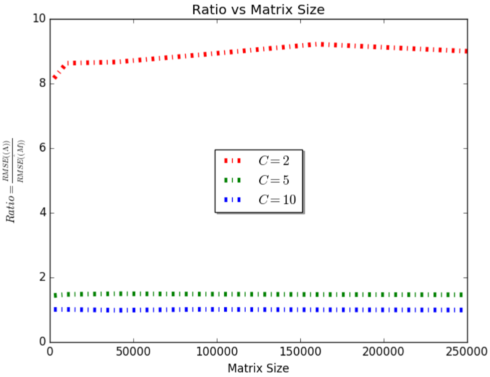

Fig. 5. Effect of increasing the size of the matrix, X^p, on the Ratio: $\frac{RMSE(\hat{\Lambda})}{RMSE(\hat{\boldsymbol{M}})}$. The different plots represent different levels of censoring. $J = 5$. The three levels of censoring are: little ($C = 10$), mild ($C = 5$) and significant ($C = 2$).

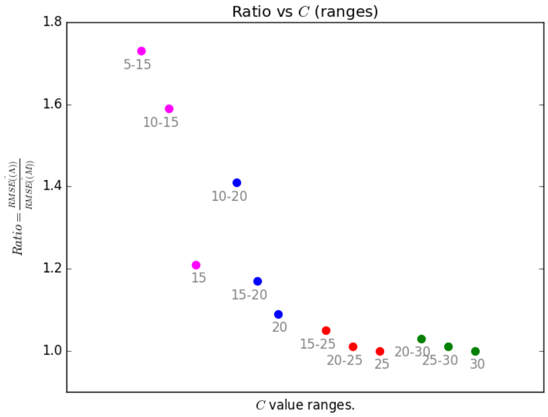

Fig. 6. Ratio vs various sets of values of C_{ij}. We have $J = 15$. The ranges of values of C_{ij} are grouped in four sets. Each set is colored differently. For instance, the red dots indicate C_{ij} values in the ranges $15 - 25$, $20 - 25$ and 25 (constant).

6.4 Effects of range of values of C_{ij}

Our simulations, thus far, have assumed a constant value for all C_{ij} to illustrate the effects of censoring. However, each location i and time period j can experience varying levels of censoring. To study the effects of censoring across ranges of of values, we assign values of stock levels to each C_{ij} within a range. We vary the range across individual experiments to study the effect on the Ratio. We expect that the larger the range, given the same upper limit, the larger the Ratio to be.

Figure 6 confirms our intuitive expectations. We have $J = 15$. Just like previously argued, for the constant values of $C_{ij} \forall i, j$, we see a drop in Ratio as the censoring effect is reduced. More interestingly, across each set of ranges of the values of C_{ij}, the Ratio is highest when more variation is allowed and drops down when the range becomes a constant. This helps us anticipate that if we know the stocking levels vary greatly across locations and time then we can expect a loss of precision in estimating the true parameters, as one might expect. In other words, more *structure*, i.e. constant $C_{ij}, \forall i, j$, leads to more precise estimates.

For completeness, we observe the same effects if we allow the scaling constants to vary randomly for each store i and time j. The larger the variation among the scaling constants, the less structure there is to exploit, and that leads to worse estimates.

6.5 The effect of structure

The premise of our work is that the imposition of some *structure* allows us to estimate the *true* demand parameters better. We confirm that by comparing the case of allowing the parameters, λ_{ij}, to be chosen randomly to the case of choosing λ_{ij} in manner outlined in earlier in Section 6.1. Figure 7 confirms that our premise is sound by demonstrating that the imposition of *structure* allows better estimation of the *true* parameters.

7 A CASE STUDY: WALMART

After extensive simulated experiments, we turn our attention to a real-world dataset. Our goal is to use actual sales data from several stores across time for a product or type of products and learn the *true* parameters of demand for each location and instance of time. To that end, we use the Walmart sales data made available by Kaggle [1]. The dataset provides sales data for 45 stores located across

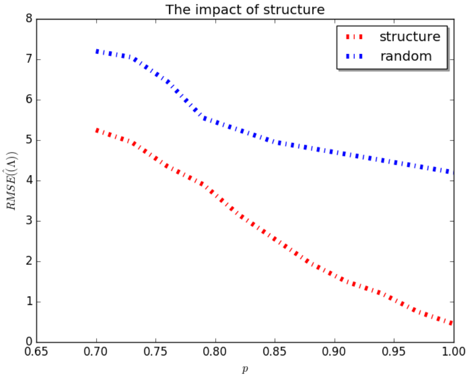

Fig. 7. $RMSE(\hat{\Lambda})$ vs p for the Poisson parameters being chosen randomly and with the structure imposed in our model. $J = 15$ and $C_{ij} = 30$ for all i and j.

different geographical regions. Each store provides weekly sales data for up to 100 departments for 143 weeks (Feb 5, 2010 - Oct 26, 2012). As such, consider the sales data for each department to be a 45 x 143 matrix of observations. Several department matrices have missing data/information.

As is typical for real-world settings, we are unaware of the *true* demand generating distributions and stocking levels at each store location and instance of time. This information is not provided with the dataset either. Therefore, we have no definitive way to evaluate how our approach performs in determining the true demand function parameters. To this end, we make certain assumptions and adopt heuristics to determine the value of this exercise.

7.1 Modeling Assumptions

For a given department, we have a 45 x 143 matrix of observations, Y. For simplicity, we firstly assume that there is little to no censoring, i.e. $C_{ij} >> Y_{ij}$ for all i, j. We relax this condition later to study the impacts of induced (artificial) censoring. Note that we assume $Y_{ij} \sim Poisson(\lambda_{ij})$. We choose a probability of observation, $p \in (0, 1]$, which results in an observation matrix X^p. This allows us to learn the parameters λ_{ij} as detailed in Section 3.

7.2 Learning parameters via De-noising

We use the observation matrix, X^p, from the Walmart sales dataset to determine the parameters of the Poisson distributions, Y_{ij}. As mentioned earlier, we do not know whether the *true* demand distributions are Poisson. However, we use our estimates $\hat{\lambda}_{ij}$ to evaluate the error between them and the actual demand observations, Y_{ij}.

Figure 8 shows the RMSE between the estimated means, $\hat{\lambda}_{ij}$ and the actual observations Y_{ij}. We vary the independent observation probability (horizontal axis) to see the effect on the RMSE. The plot shows the RMSE computed across all the entries of the matrix and also just for the hidden entries. It is clear that RMSE values are lower as p gets higher. Given that the RMSE values are similar for both the entire matrix and for the values that were hidden, there appears to be structure in the data which has been exploited by our method. We call this property a **de-noising** effect because on average our estimated means of the true demand are not too far off from the observations, on average.

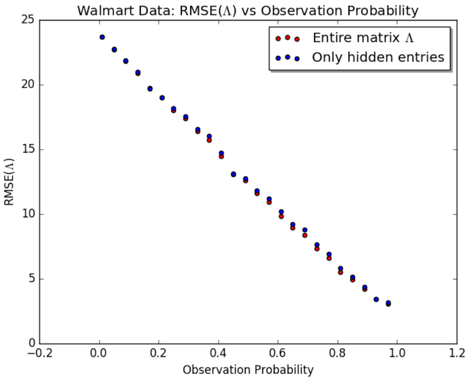

Fig. 8. For the Walmart sales data across 45 stores and 143 weeks. Department = 79. The plot shows RMSE($\hat{\Lambda}$) vs observation probability, p. The RMSE is obtained between the estimated $\hat{\lambda}_{ij}$ and the original observaions Y. We assume little to no-censoring. The plot is showing RMSE values for the entire matrix $\hat{\Lambda}$ and also for only those values that were hidden (due to our choice of p).

7.3 Transformed Distribution of Observations: Gaussian

Thus far, we have established that our approach allows us to approximate the de-noising of the data observations reasonably well, on average. However, we do not know the *true* demand distributions. Therefore, one natural question is to evaluate how valid our model assumption about the demand being a matrix of independent Poisson variables with means λ_{ij} really is. To that end, we use the bootstrap method to generate the distribution of the following random variable:

$$\mathbb{W} = \frac{\frac{1}{|\mathbb{S}|}\sum\limits_{ij\in\mathbb{S}}(Y_{ij}-\hat{\lambda_{ij}})}{\sqrt{\frac{1}{(|\mathbb{S}|)^2}\sum\limits_{ij\in\mathbb{S}}\hat{\lambda_{ij}}}}$$

where $\mathbb{S}$ is a random sample of the indices of observation matrix Y and the estimated mean matrix $\hat{\Lambda}$.

If the entries in the matrix Y are indeed independent Poissons, we expect $\mathbb{W}$ to be Normally distributed because for a Poisson random variable with mean λ, the variance is also λ. Figure 9 shows that the histogram and the QQ-plot both show an approximately Normal distribution of $\mathbb{W}$. Both plots confirm a center to the right of 0 which suggests that there is some censoring in the dataset (see Section 7.4). The QQ-plot shows that the data points lie on the red straight line which confirms Normality with reference to a Gaussian distribution with mean equal to the sample mean and standard deviation equation to the sample mean's standard deviation. Remarkably, this appears to suggest that our assumption about the data being distributed as independent Poisson random variables is the valid for this dataset.

7.4 Estimated parameters as Lower bounds

Notice that while Figure 8 shows the RMSE decaying, it doesn't reach zero. The estimated parameters, $\hat{\lambda}_{ij}$, tend to be lower-bounds of the observations. Figure 10 shows this behavior via a comparison of the average of observations, $\frac{1}{NT}\sum\limits_{i,j} Y_{ij}$, and the average of the estimated means, $\frac{1}{NT}\sum\limits_{i,j}\hat{\lambda}_{ij}$. This plot confirms the findings in the RMSE plot in Figure 8 where the average of the estimated means approaches the average of the observations as p increases. However, the estimated averages are always a lower-bound on the averages of the observations. We find that this lower-bound behavior

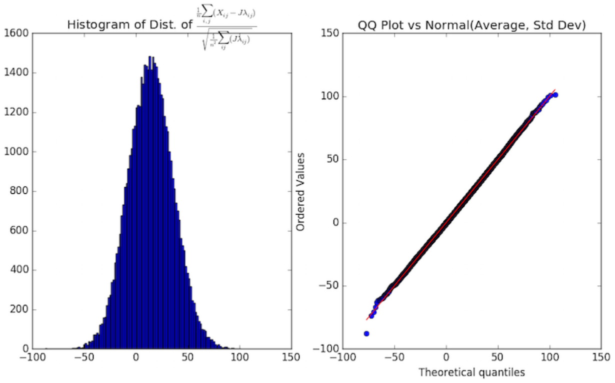

Fig. 9. For the Walmart sales data across 45 stores and 143 weeks. Department = 79. The histogram on the left is generated by random sampling from the random variable $\mathbb{W}$. The matrices are sampled at random with an independent probability of selection, i.e. $p < 1$. On the the right is a QQ-plot 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 $\mathbb{W}$.

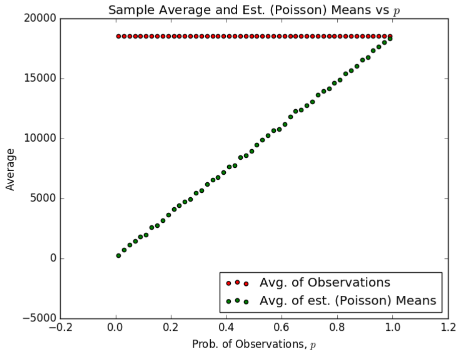

Fig. 10. For the Walmart sales data across 45 stores and 143 weeks. Department = 79. The plot shows the comparison between the average of observations matrix and the average of the estimated means. We assume no censoring. The observation probability, p, is varied (horizontal axis).

holds across all departments in the Walmart dataset. This finding is useful because it hints at the utility of this approach in planning exercises for retailers where conservative estimates of the demand functions can be made by following the approach introduced in this paper. Note also that the estimates could be improved with knowledge of the actual stock levels, C_{ij}, which were assumed to be practically infinite in this series of experiments.

7.5 Induced Censoring

Given that we have established that our assumption about the demand data being Poisson is reasonable for the Walmart data, we next investigate whether the original data parameters can be learned after some induced artificial censoring. We censor the observations by choosing a stocking level, C_{ij}, which is not as large as the one chosen in the experiments described earlier. Eventually, we learn the parameters $\hat{\lambda}_{ij}$ as estimates of the true demand distributions.

We choose mild and significant censoring. For the mild censoring situation we set $C_{ij} = 0.4(\max_{i.j} Y_{ij})$ and significant censoring where we set $C_{ij} = 0.2(\max_{i.j} Y_{ij})$. The stocking levels, C_{ij}, are all set to

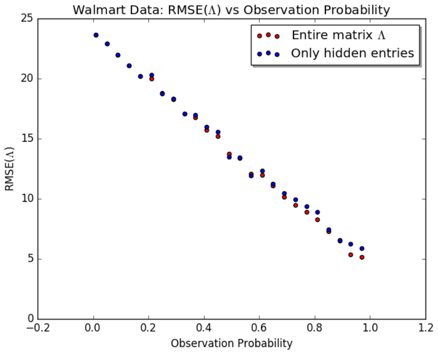

Fig. 11. Mild induced censoring. For the Walmart sales data across 45 stores and 143 weeks. Department = 79. The plot shows RMSE($\hat{\Lambda}$) vs p. The RMSE is obtained between the estimated $\hat{\Lambda}$ and the original observations Y. We keep $C_{ij} = 0.4(\max_{i,j} Y_{ij})$ for all i, j. The plots show the RMSE values for the entire matrix $\hat{\Lambda}$ and also for only those values that were hidden (due to our choice of p).

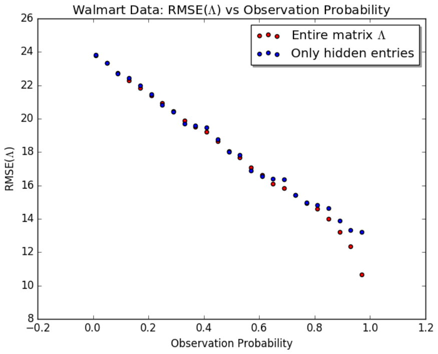

Fig. 12. Significant induced censoring. For the Walmart sales data across 45 stores and 143 weeks. Department = 79. The plot shows RMSE($\hat{\Lambda}$) vs p. The RMSE is obtained between the estimated $\hat{\Lambda}$ and the original observations Y. We keep $C_{ij} = 0.2(\max_{i,j} Y_{ij})$ for all i, j. The plots show the RMSE values for the entire matrix $\hat{\Lambda}$ and also for only those values that were hidden (due to our choice of p).

the same value within each experiment. We note that the mild censoring situation ends up censoring about 30% of the entries in the original Walmart dataset, department 79. In the significant censoring case we notice about 66% of the entries experiencing censoring. Figures 11 (mild-censoring) and 12 (significant censoring) show the plots obtained for $RMSE(\hat{\Lambda})$ with reference to the Walmart data observations. Compare these plots to Figure 8, while noticing the scale differences on the vertical axis, which shows the same plots for the situation with no censoring. As the amount of (induced) censoring is increased the RMSE values increase confirming our intuition from the simulated experiments that the estimates get worse with censoring.

8 DISCUSSION

Estimation of *true* demand parameters from noisy, incomplete and censored sales data is a problem of significant interest. We present a novel approach to estimating the *true* demand parameters from a *single* sample of a matrix of observations across N stores and for T time periods. We assume that the demand at each location and time period is distributed as a Poisson random variable. We model the demand with independent, but not identical, Poisson random variables in a latent variable setting. This allows us to present a spectral algorithm to estimate the *true* parameters from the matrix of observations. Note that our approach is not restricted to Poisson random variables. See Appendix A for an example with a different distribution. However, if a practitioner's assumption about the true demand distribution is incorrect, the procedure suggested here will produce an error quantifying which is not the subject of this work. We show that our estimates for the censored means and the *true* demand parameters are consistent, i.e. the average expected $MSE \longrightarrow 0$ as $N, T \longrightarrow \infty$. Further, we show that as the degree of censoring increases the estimates become poorer and establish this analytically and with the help of simulations. Finally, we conduct a series of experiments on a real-world dataset with Walmart's sales data and conclude that our approach has great practical value.

ACKNOWLEDGMENTS

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

REFERENCES

[1] 2014. (2014). https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

[2] Oren Anava, Elad Hazan, and Assaf Zeevi. 2015. Online Time Series Prediction with Missing Data. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, David Blei and Francis Bach (Eds.). JMLR Workshop and Conference Proceedings, 2191–2199. http://jmlr.org/proceedings/papers/v37/anava15.pdf

[3] Ravi Anupindi, Maqbool Dada, and Sachin Gupta. 1998. Estimation of Consumer Demand with Stock-Out Based Substitution: An Application to Vending Machine Products. *Marketing Science* 17, 4 (1998), 406–423. DOI:http://dx.doi.org/10.1287/mksc.17.4.406 arXiv:http://dx.doi.org/10.1287/mksc.17.4.406

[4] Katy S. Azoury. 1985. Bayes Solution to Dynamic Inventory Models Under Unknown Demand Distribution. *Management Science* 31, 9 (2017/01/07 1985), 1150–1160. DOI:http://dx.doi.org/10.1287/mnsc.31.9.1150

[5] Gah-Yi Ban. 2015. The data-driven (s, S) policy: The data-driven (s, S) policy: The data driven (s, S) policy: why you can have confidence in censored demand data. *Available at SSRN: https://ssrn.com/abstract=2654014* (2015).

[6] Apostolos N. Burnetas and Craig E. Smith. 2000. Adaptive Ordering and Pricing for Perishable Products. *Operations Research* 48, 3 (2017/01/07 2000), 436–443. DOI:http://dx.doi.org/10.1287/opre.48.3.436.12437

[7] Sourav Chatterjee. 2015. Matrix estimation by Universal Singular Value Thresholding. *Ann. Statist.* 43 (2015), 177–214.

[8] Li Chen and Adam J. Mersereau. 2015. *Analytics for Operational Visibility in the Retail Store: The Cases of Censored Demand and Inventory Record Inaccuracy.* Springer US, Boston, MA, 79–112. DOI:http://dx.doi.org/10.1007/978-1-4899-7562-1_5

[9] Zhe Chen and Andrzej Cichocki. 2005. Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints. *Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep* 68 (2005).

[10] Christopher T. Conlon and Julie Holland Mortimer. 2013. Demand Estimation under Incomplete Product Availability. *American Economic Journal: Microeconomics* 5, 4 (November 2013), 1–30. DOI:http://dx.doi.org/10.1257/mic.5.4.1

[11] S. A. Conrad. 1976. Sales Data and the Estimation of Demand. *Operational Research Quarterly (1970-1977)* 27, 1 (1976), 123–127. http://www.jstor.org/stable/3009217

[12] Gregory A. Godfrey and Warren B. Powell. 2001. An Adaptive, Distribution-Free Algorithm for the Newsvendor Problem with Censored Demands, with Applications to Inventory and Distribution. *Management Science* 47, 8 (2017/01/07 2001), 1101–1112. DOI:http://dx.doi.org/10.1287/mnsc.47.8.1101.10231

[13] Woonghee Tim Huh and Paat Rusmevichientong. 2009. A Nonparametric Asymptotic Analysis of Inventory Planning with Censored Demand. *Mathematics of Operations Research* 34, 1 (2017/01/07 2009), 103–123. DOI:http://dx.doi.org/10.1287/moor.1080.0355

[14] Sumit Kunnumkal and Huseyin Topaloglu. 2008. Using Stochastic Approximation Methods to Compute Optimal Base-Stock Levels in Inventory Control Problems. *Operations Research* 56, 3 (2017/01/07 2008), 646–664. DOI:http://dx.doi.org/10.1287/opre.1070.0477

[15] Retsef Levi, Robin O. Roundy, and David B. Shmoys. 2007. Provably Near-Optimal Sampling-Based Policies for Stochastic Inventory Control Models. *Mathematics of Operations Research* 32, 4 (2017/01/07 2007), 821–839. DOI:http://dx.doi.org/10.1287/moor.1070.0272

[16] L. Massoulié M. Lelarge and J. Xu. 2014. Edge label inference in generalized stochastic block model: from spectral theory to impossibility results. *Conference on Learning Theory (COLT)* (2014).

[17] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research* 11, Aug (2010), 2287–2322.

[18] Andrés Musalem, Marcelo Olivares, Eric T. Bradlow, Christian Terwiesch, and Daniel Corsten. 2010. Structural Estimation of the Effect of Out-of-Stocks. *Management Science* 56, 7 (2017/01/07 2010), 1180–1197. DOI:http://dx.doi.org/10.1287/mnsc.1100.1170

[19] Steven Nahmias. 1994. Demand estimation in lost sales inventory systems. *Naval Research Logistics (NRL)* 41, 6 (1994), 739–757. DOI:http://dx.doi.org/10.1002/1520-6750(199410)41:6<739::AID-NAV3220410605>3.0.CO;2-A

[20] Warren Powell, Andrzej Ruszczyński, and Huseyin Topaloglu. 2004. Learning Algorithms for Separable Approximations of Discrete Stochastic Optimization Problems. *Mathematics of Operations Research* 29, 4 (2017/01/07 2004), 814–836. DOI:http://dx.doi.org/10.1287/moor.1040.0107

[21] Swati Rallapalli, Lili Qiu, Yin Zhang, and Yi-Chao Chen. 2010. Exploiting Temporal Stability and Low-rank Structure for Localization in Mobile Networks. In *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking (MobiCom '10)*. ACM, New York, NY, USA, 161–172. DOI:http://dx.doi.org/10.1145/1859995.1860015

[22] Matthew Roughan, Yin Zhang, Walter Willinger, and Lili Qiu. 2012. Spatio-temporal Compressive Sensing and Internet Traffic Matrices. *IEEE/ACM Trans. Netw.* 20, 3 (June 2012), 662–676. DOI: http://dx.doi.org/10.1109/TNET.2011.2169424

[23] Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic matrix factorization. In *NIPS*, Vol. 20. 1–8.

[24] Ruslan Salakhutdinov and Andriy Mnih. 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*. ACM, 880–887.

[25] Catalina Stefanescu. 2009. Multivariate Customer Demand: Modeling and Estimation from Censored Sales. *Available at SSRN: https://ssrn.com/abstract=1334353* (2009).

[26] Gustavo Vulcano, Garrett van Ryzin, and Richard Ratliff. 2012. Estimating Primary Demand for Substitutable Products from Sales Transaction Data. *Operations Research* 60, 2 (2012), 313–334. DOI: http://dx.doi.org/10.1287/opre.1110.1012 arXiv:http://dx.doi.org/10.1287/opre.1110.1012

[27] William E. Wecker. 1978. Predicting Demand from Sales Data in the Presence of Stockouts. *Management Science* 24, 10 (1978), 1043–1054. http://www.jstor.org/stable/2630558

[28] Liang Xiong, Xi Chen, Tzu-Kuo Huang, Jeff Schneider, and Jaime G. Carbonell. 2010. *Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization.* 211–222. DOI: http://dx.doi.org/10.1137/1.9781611972801.19 arXiv:http://epubs.siam.org/doi/pdf/10.1137/1.9781611972801.19

[29] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S. Dhillon. 2015. Temporal Regularized Matrix Factorization. *CoRR* abs/1509.08333 (2015).

Appendices

A OTHER PROBABILITY DISTRIBUTIONS

As discussed earlier, our setting is not limited to situations where the demand is distributed as Poisson. While Poisson-distributed demand is natural for the retail setting, other applications may warrant data distributed differently. While different distributions will lead to different forms and constants in the results derived earlier, our approach will extend to any probability distribution that allows the following to hold:

(i). **Censored mean as a function of parameters and C.** As discussd in Section 5.1.1, the censored mean, m_{ij} must be a function of the parameters of the distribution and the stocking levels, C_{ij}. We also require that this function be continuous and differentiable in all parameters, given C_{ij}.

(ii). **Parameters as functions of location and time.** We require that the parameters be functions of hidden variables of time, θ_i, and location, ρ_j. In other others, if the d-dimensional parameters are represented by λ_{ij}^d, then we must have $\lambda_{ij}^d = \boldsymbol{h}(\theta_i, \rho_j)$. For a direct application of Lemma 4.1, we require that $\boldsymbol{h}(\theta_i, \rho_j)$ be Lipschitz and $\mathbb{P}(C_{ij} = k)$ also be a Lipschitz function of θ_i, ρ_j.

As an example of a distribution other than Poisson, consider that the demand at each location i and time period j is distribution as a Binomial distribution with a known parameter n_{ij} and unknown parameter λ_{ij}. We have that $Y_{ij} \sim Binomial(n_{ij}, \lambda_{ij})$. As usual, we define $X_{ij} = \min\{Y_{ij}, C_{ij}\}$ and each X_{ij} is observed with probability $p \in (0, 1]$. The censored means are defined as $m_{ij} = \mathbb{E}(X_{ij})$. It is straight forward to show that, ignoring the subscripts for simplicity, $m = \boldsymbol{f}(\lambda, n, C)$, noting that we have assumed that n is a known (scaling) constant.

$$\begin{aligned} m &\equiv \mathbb{E}[X] \\ &= \sum_{k=0}^{C-1} k\mathbb{P}(Y = k) + C\big(\sum_{k=C}^{\infty} \mathbb{P}(Y = k)\big) \\ &= \sum_{k=0}^{\infty} k\mathbb{P}(Y = k) - \sum_{k=C}^{\infty} (k - C)\mathbb{P}(Y = k) \\ &= \mathbb{E}[Y] - \sum_{k=C}^{\infty} (k - C)\mathbb{P}(Y = k) \\ &= n\lambda - \sum_{k=C}^{\infty} (k - C)\binom{n}{k}\lambda^k(1 - \lambda)^{(n-k)} \\ &\equiv f(\lambda, n, C). \end{aligned} \tag{49}$$

We can see that $\boldsymbol{f}(.)$ is continuous in the parameter λ_{ij}, given C_{ij} and n_{ij}. This satisfies the first assumption stated above. If the conditions stated in the second assumptions hold then similar to the Poisson case we can show that $\boldsymbol{f}(\theta_i, \rho_j)$ is Lipschitz continuous in the latent parameters. The algorithm described in Section 3 can then be used to compute the estimates of censored means, $\hat{m}_{ij}$, and original parameters, $\hat{\lambda}_{ij}$. In this case, the results stated in Section 4 are directly applicable.

B USVT VS GSBM ALGORITHM

We choose the USVT Algorithm of [7] for *Step 1* of our Algorithm. The USVT algorithm is a spectral thresholding algorithm and there are several competing spectral algorithms in literature which could also be considered. However, the USVT algorithm is attractive due to its *universal* nature–it proposes a threshold whereas several others in literature do not prescribe one, i.e. they are existential in nature. Secondly, the USVT algorithm is applicable to symmetric *and* non-symmetric matrices whereas several other algorithms in literature deal specifically with symmetric ones. Finally, the USVT is a computationally efficient algorithm within the class of spectral methods.

For comparison, we consider the Generalized Stochastic Block Model (GSBM) algorithm proposed in [16]. Since this algorithm is only applicable to symmetric matrices, our comparisons with the USVT algorithm consider only symmetric matrices. We consider the following setting:

We generate N parameters, θ, and keep them fixed. In this case $\theta_i \sim Uniform(0, 1), 1 \leq i \leq N$. We let $\alpha \sim Uniform(0, 1)$. Next, we generate the following probabilities of edges between each pair of vertices:

$$p_{ij} = \frac{1}{1 + \exp\{f_{ij}\}}, 1 \leq i, j \leq N,$$

where

$$f_{ij} = \exp\{-\theta_i - \theta_j - \alpha\theta_i\theta_j\}$$

Edges, $e_{ij} \in \{0, 1\}$ are then generated for the graph using the interaction (edge) probabilities p_{ij}. Further, each realization of an edge is observed with probability p. The generated realization of edges is our graph, G.

In our implementation, we set the following GSBM algorithm specific parameters [16]:

$$\epsilon = \frac{1}{2}\text{median}\{||z_i - z_j||\}$$

and,

$$h_\epsilon(x) = \min\{1, \max\{0, 2 - \frac{x}{\epsilon}\}\}$$

The number of eigenvalues retained is kept to be the same as the number chosen by the USVT algorithm since [16] does not provide a guideline for choosing how many to retain. The only USVT algorithm specific parameter [7] is $\eta = 0.01$.

We conduct two types of comparison experiments: (a) fix the (symmetric) dimension, N, of the matrix, G, and vary the probability of observation, p; (b) keep the p fixed and vary the dimension of the matrix. In each case, the metric of interest is the average $MSE(\hat{G})$ with respect to G. Figures 13 and 14 show the average MSE produced by each algorithm for the experiments (a) and (b). We conclude that the USVT algorithm is a better choice. It is also more versatile (applicable to non-symmetric settings), prescriptive (specifies exactly what threshold to choose) and is computationally far superior.

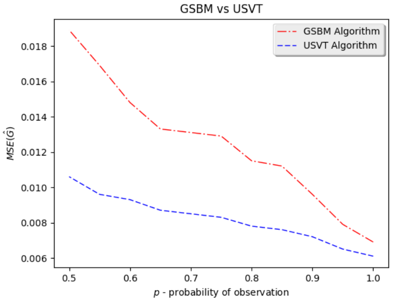

Fig. 13. Average *MSE* vs the probability of observation for the two algorithms under consideration. The size of matrix is fixed to $80x80$. Note that the USVT algorithm is comparatively better at all p.

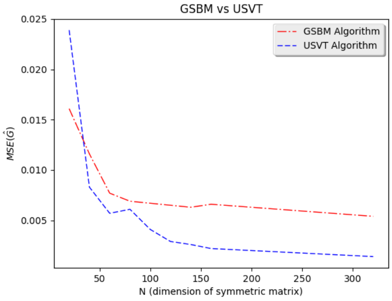

Fig. 14. Average *MSE* vs the size of the (symmetric) matrix for the two algorithms under consideration. The probability of observation, $p = 1.0$. Note that the USVT algorithm performs comparatively better at all but the smallest matrix size.