Solving the Phantom Inventory Problem: Near-optimal Entry-wise Anomaly Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

We observe that a crucial inventory management problem ('phantom inventory'), that by some measures costs retailers approximately 4% in annual sales can be viewed as a problem of identifying anomalies in a (low-rank) Poisson matrix. State of the art approaches to anomaly detection in low-rank matrices apparently fall short. Specifically, from a theoretical perspective, recovery guarantees for these approaches require that non-anomalous entries be observed with vanishingly small noise (which is clearly not the case in our problem, and indeed in many applications). So motivated, we propose a conceptually simple entry-wise approach to anomaly detection in low-rank Poisson matrices. Our approach accommodates a general class of probabilistic anomaly models. We extend recent work on entry-wise error guarantees for matrix completion, establishing such guarantees for sub-exponential matrices, where in addition to missing entries, a fraction of entries are corrupted by (an also unknown) anomaly model. We show that for any given budget on the false positive rate (FPR) our approach achieves a TPR that approaches the TPR of an optimal algorithm at a min-max optimal rate. Using data from a massive consumer goods retailer, we show that our approach provides significant improvements over incumbent approaches to anomaly detection.

1 Introduction

Consider the problem of identifying anomalies in a low-rank matrix: Specifically, let M^* be some low-rank matrix, and let Y be a random matrix with independent entries and expected value M^*. Finally, let $X = Y + A$ where A is an unknown, sparse, matrix of anomalies. We observe X on some subset of matrix entries Ω. The anomaly detection problem concerns identifying the support of A simply from these observations. State-of-the-art approaches to solving this problem stem from algorithms for matrix completion; for instance, consider solving the following convex optimization problem (referred to as 'Stable PCP' [1]) where λ_1 and λ_2 are regularization parameters:

$$\min_{\hat{Y},\hat{A}} \|\hat{Y}\|_*^2 + \lambda_2\|\hat{A}\|_1 + \lambda_1\|P_\Omega(X - \hat{Y} - \hat{A})\|_F^2 \tag{1}$$

We may then use $\hat{A}$ to identify the support of A. Now in the absence of anomalies, the optimization problem above (after removing the $\hat{A}$ terms) is, in essence optimal under a variety on assumptions on the distributions of Y and Ω. In contrast, the available results for anomaly detection are weaker. Perhaps most limiting, results that guarantee the recovery of A, require the total observation noise $\|Y - M^*\|_F$ be bounded by a constant independent of the size of the matrix. In this setting, noise in observing any individual matrix entry in Ω grows negligibly small in large matrices. This is limiting:

1. Y is typically noisy: In the practical problem that motivates this work, Y can be viewed as a matrix of Poisson entries with mean M^*. Clearly then, $\mathbb{E}\|Y - M^*\|_F$ will scale with the

size of the matrix so theoretical guarantees for extant anomaly detection approaches do not apply.

2. Even ignoring this theoretical limitation, we will see that in the setting where Y is noisy (such as in our motivating practical problem), the optimization approach above can perform quite poorly.

This Paper: Against the above backdrop, we develop a new anomaly detection algorithm for low-rank Poisson matrices. Under a broad class of probabilistic anomaly models we prove that our approach is min-max optimal for this problem.

Our results are powered by two ingredients. First, we generalize recent entry-wise guarantees for matrix completion to sub-exponential matrices. Next, we show that combined with a moment matching approach to learning the anomaly model, we can jointly learn the anomaly model *along with* the true underlying rate matrix. We obtain entry-wise guarantees here that match those one would obtain absent anomalies. This in turn suffices to build a classification algorithm that we show is near optimal in the sense that it achieves an ROC curve that converges to the optimal ROC curve at a min-max optimal rate. The min-max optimality is established through a hypothesis testing argument.

Our work is motivated by a crucial inventory management problem ('phantom inventory'; a thorough description is deferred to later) that costs the retail industry up to 4% in annual revenue. We observe that this inventory problem can be viewed as one of detecting anomalies in a low-rank Poisson matrix. The latter is the matrix one obtains by viewing sales data in matrix-form with rows corresponding to store locations, columns corresponding to products, and entries corresponding to observed sales over some, typically short, period – say, a week. On large-scale data (thousands of stores, thousands of products) we find that our approach significantly outperforms the convex optimization approach to detecting anomalies.

Related Literature: There are three ongoing streams of work to which the present paper contributes. The first, naturally, is in anomaly detection for matrices. The majority of these studies has focused on a formulation called *robust principal component analysis (PCA)* [2, 3], and in particular, approaches based on convex relaxations. Most relevant to our problem (which allows for noise) is the *stable PCP* [1], written in Eq. (1). Despite a sequence of breakthroughs and improvements in algorithms for optimizing these convex objectives (initial work by [4, 5, 6]; see [7, 8] for surveys of more recent work), progress in statistical guarantees for these formulations has been relatively slower since the initial results of [2, 1]. Recent progress has been on more refined observation models [9] and on guarantees for nonconvex objectives [10, 11, 12]. Still, the overall state as it pertains to the model we will propose next is limited to the status presented in the introduction: the relevant existing guarantees will be insufficient, and for good reason – these models effectively allow for *adversarial* perturbations, so accuracy is naturally limited.

The second body of work concerns statistical inference in matrix completion. This stream [13, 14, 15] has recently produced tight statistical characterizations of various algorithms for random matrices. Our own algorithm necessitates proving a similar result, borrowing crucial techniques from [13] in particular to prove the first entry-wise guarantee for sub-exponential (rather than sub-gaussian) noise. Adjacent to this stream is work on matrix completion with Poisson observations in particular [16], from which we also borrow.

Finally, with respect to our motivating application: phantom inventory is well-studied in the area of Operations Management. The phenomenon itself has been observed for some time [17, 18], with observed causes ranging from theft [19], to misplacement [20], to point-of-sale errors [21]. Despite technological progress in inventory tracking, phantom inventory remains a primary challenge for retailers [22]. Existing algorithmic solutions have focused on [23, 24] adapting inventory management policies to uncertain inventory levels. Algorithmic *detection*, particularly in a form that combines observations across products and stores, is the motivation for this work.

Notation: The sub-exponential norm of X is defined as $\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}\left(\exp(|X|/t)\right) \leq 2\}$. For $A \in \mathbb{R}^{n\times m}$, we write $\sum_{(i,j)\in[n]\times[m]} A_{ij}$ as $\sum_{ij} A_{ij}$ when no ambiguity exists. $\|A\|_{2,\infty} := \max_i \sqrt{\sum_j A_{ij}^2}$, $\|A\|_{\max} = \max_{ij} |A_{ij}|$, $\|A\|_{\mathrm{F}} = \sqrt{\sum_{ij} A_{ij}^2}$. The letter C (and c) represents a sufficiently large (and small) universal (i.e. not dependent on problem parameters) constant that may change between equations.

2 Model

We are given (an unobserved) 'rate' matrix $M^* \in \mathbb{R}_+^{n\times m}$ ($n \le m$ without loss of generality). A second matrix $B \in \{0,1\}^{n\times m}$ serves to indicate the position of anomalies. Given M^* and B, we generate a *random* matrix X with independent entries distributed according to

$$X_{ij} \sim \begin{cases} \text{Poisson}(M^*_{ij}) & \text{if } B_{ij} = 0 \\ \text{Anom}(\alpha^*, M^*_{ij}) & \text{if } B_{ij} = 1. \end{cases}$$

$\text{Anom}(\cdot,\cdot)$ is some non-negative, integer-valued random variable and $\alpha^* \in \mathbb{R}^d$ is an unknown parameter vector. We *observe* X_Ω where $\Omega \subset [n] \times [m]$ is random. Specifically, we assume that entries are observed independently with probability p_{O}. In addition, we assume that B is a $\text{Bernoulli}(p^*_{\mathrm{A}})$ random matrix where p^*_{A} is bounded away from one by a constant.

Our goal is to infer B given X_Ω. We discuss next how this model fits the phantom inventory problem, and the assumptions we place on M^* and the anomaly distribtution.

Fit to Application: In the Phantom inventory problem, X is a sales matrix so that the (i,j)th entry corresponds to sales of product j at store i; the Poisson distribution is typically a good fit for sales data [25, 26]. Our results will not rely on the Poisson assumption; any sub-exponential, integer-valued random variable will do. Anomalies in this setting are the consequence of so-called shelf-execution errors and typically result in a censoring of sales so that for our motivating problem $\text{Anom}(\alpha^*, \lambda)$ is perhaps best viewed as a censored $\text{Poisson}(\lambda)$ random variable. Again, our results will allow for a broad family of distributions for anomalies, which we describe momentarily.

Assumptions on M^*: Let $M^* = U\Sigma V^T$, be the SVD of M^*, where $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix with singular values $\sigma_1^* \ge \sigma_2^* \ge \ldots \ge \sigma_r^*$ ($\kappa = \sigma_1^*/\sigma_r^*$); and $U \in \mathbb{R}^{n\times r}, V \in \mathbb{R}^{m\times r}$ are two matrices that hold the left and right-singular vectors. We make the following assumptions:

- (Boundedness): $\|M^*\|_{\max} + 1 \le L$.
- (Incoherence): $\|U\|_{2,\infty}^2 \le \frac{\mu r}{n}, \|V\|_{2,\infty}^2 \le \frac{\mu r}{m}$.
- (Sparsity): $p_{\mathrm{O}} \ge \max\left(C_1 \frac{n^2}{\sum_{ij} M^*_{ij} \log m}, C_2 \frac{\sqrt{m} L \log(m) \kappa^2 \mu r}{n}\right)$ for some constant C_1 and a known constant C_2.

Similar to existing results in matrix completion and recovery, our guarantees will be parameterized by μ, L, r and κ.

Assumptions on $\text{Anom}(\cdot,\cdot)$: For any $M = M^*_{ij}$, we have the following assumptions:

- (Sub-exponential): $\text{Anom}(\alpha^*, M)$ is sub-exponential: $\|\text{Anom}(\alpha^*, M)\|_{\psi_1} \le L$.
- (Lipschitz): For all $k \in \mathbb{N}$, $\mathbb{P}(\text{Anom}(\alpha^*, M) = k)$ is K-Lipschitz in (α^*, M).
- (Mean Decomposition): We have $\mathbb{E}(\text{Anom}(\alpha^*, M)) = g(\alpha^*)M$ for some $g : \mathbb{R}^d \to \mathbb{R}$.

We pause to discuss the restrictiveness of our assumptions. To begin, we assume a probabilistic anomaly model as opposed to one that is adversarial. Insomuch as our requirements of this model are concerned the mean decomposition is perhaps the most restrictive. Nonetheless, we suspect that these assumptions are parsimonious enough to allow for many practical applications: in our motivating application, this assumption is well justified from the known mechanisms for anomalies (essentially, random censoring of sales). It is also worth considering that alternative anomaly detection models that allow for adversarial anomalies require in essence *exact* observations of M^* when the observations are non-anomalous. This is highly problematic in many applications, including our motivating application – sales in absence of anomalies will be highly noisy, especially over a short period of time.

2.1 Performance Metrics

Let $A^\pi(X_\Omega)$ be some estimator of B. Given X_Ω, we define the true positive rate for this estimator, $\text{TPR}_\pi(X_\Omega)$ as the ratio of the expected number of true positives under the algorithm and the expected number of anomalies given X_Ω. We similarly define the false positive rate, $\text{FPR}_\pi(X_\Omega)$.

More formally, let f_{ij}^* be the conditional probability that the (i, j)th entry is not anomalous, given X, i.e. $f_{ij}^* := \mathbb{P}(B_{ij} = 0 \mid X)$. Then, some algebraic manipulation establishes

$$\mathrm{TPR}_\pi(X_\Omega) = \frac{\sum_{(i,j)\in\Omega} \mathbb{P}\left(A_{ij}^\pi(X) = 1\right)(1 - f_{ij}^*)}{\sum_{(i,j)\in\Omega}(1 - f_{ij}^*)} \tag{2}$$

$$\mathrm{FPR}_\pi(X_\Omega) = \frac{\sum_{(i,j)\in\Omega} \mathbb{P}\left(A_{ij}^\pi(X) = 1\right) f_{ij}^*}{\sum_{(i,j)\in\Omega} f_{ij}^*}. \tag{3}$$

Our goal will be to maximize TPR for some bound on FPR. In establishing the quality of our algorithm we will compare, for a given constraint on FPR, the TPR achieved under our algorithm to that achieved under the optimal estimator. We will show that in large matrices this gap grows negligibly small at a min-max optimal rate.

3 Algorithm and Theoretical Guarantees

We are now prepared to state our approach to the anomaly detection problem formulated above. Our algorithm, which we refer to as the *entry-wise (EW)* algorithm, leverages an entry-wise matrix completion guarantee for sub-exponential noise that we will describe shortly. Besides the observed data X_Ω, the only other input into the EW algorithm is a target FPR which we denote as γ. The full algorithm is stated in Algorithm 1 below:

Algorithm 1 Entry-wise (EW) Algorithm $\pi^{\mathrm{EW}}(\gamma)$

Input: $X_\Omega, \gamma \in (0, 1]$

1: Set $\hat{M} = \frac{nm}{|\Omega|} \arg\min_{\mathrm{rank}(M)\le r} \|M - X'\|_{\mathrm{F}}$, where X' is obtained from X_Ω by setting unobserved entries to 0.

2: Estimate $\hat{\theta} = (\hat{p}_{\mathrm{A}}, \hat{\alpha})$ based on the moment matching estimator in Eq. (4) .

3: Estimate a confidence interval $[f_{ij}^{\mathrm{L}}, f_{ij}^{\mathrm{R}}]$ for f_{ij}^* for each $(i, j) \in \Omega$ according to Eq. (5).

4: Let $\{t_{ij}^{\mathrm{EW}}\}$ be an optimal solution to the following optimization problem:

$$\mathcal{P}^{\mathrm{EW}} : \max_{\{0\le t_{ij}\le 1, (i,j)\in\Omega\}} \sum_{(i,j)\in\Omega} t_{ij}$$
$$\text{subject to} \sum_{(i,j)\in\Omega} t_{ij} f_{ij}^{\mathrm{R}} \le \gamma \sum_{(i,j)\in\Omega} t_{ij} f_{ij}^{\mathrm{L}}$$

5: For every $(i, j) \in \Omega$, generate $A_{ij} \sim \mathrm{Ber}(t_{ij}^{\mathrm{EW}})$ independently.

Output: A_Ω

The goal of the EW algorithm is to maximize the TPR subject to a FPR below the input target value of γ. Our main result is the following guarantee, which states that (a) the 'hard' constraint on the FPR is satisfied with high probability, and (b) the TPR is within an additive *regret* of a certain unachievable policy we use as a proxy for the best achievable policy. Specifically, for any $\gamma \in (0, 1]$, let $\pi^*(\gamma)$ denote the optimal policy when M^*, p_{A}^*, and α^* are known (this policy is described later in this section). One can verify that, for any γ, X_Ω and policy π, $\mathrm{TPR}_{\pi^*(\gamma)}(X_\Omega) \ge \mathrm{TPR}_\pi(X_\Omega)$ if $\mathrm{FPR}_\pi(X_\Omega) \le \gamma$. Note that the only additional assumptions we require, beyond those stated in Section 2, are the set of regularity conditions (RC) stated later in this section.

Theorem 1. *Assume that the regularity conditions* (RC) *hold. Then for any* $0 < \gamma \le 1$*, with probability* $1 - \frac{1}{nm}$,

$$\mathrm{FPR}_{\pi^{\mathrm{EW}}(\gamma)}(X_\Omega) \le \gamma,$$
$$\mathrm{TPR}_{\pi^{\mathrm{EW}}(\gamma)}(X_\Omega) \ge \mathrm{TPR}_{\pi^*(\gamma)}(X_\Omega) - C\frac{(K+L)^2 L^2 \kappa^4 \mu r}{p_{\mathrm{O}}^2 p_{\mathrm{A}}^* \gamma} \frac{\log(m)\sqrt{m}}{n}.$$

In a typical application, we can expect the problem parameters to fall in the following scaling regime: $K, L, \kappa, r, \mu = O(1)$, $p_{\mathrm{O}}, p_{\mathrm{A}}^*, \gamma = \Omega(1)$, and $m/n = \Theta(1)$. For this regime, the regret is $O\left(\frac{\log n}{\sqrt{n}}\right)$, which we will see in Section 3.2 is optimal up to logarithmic factor.

3.1 Algorithm Details and Proof Sketch

In the remainder of this subsection, we motivate the steps of Algorithm 1 and provide a proof sketch of Theorem 1. Mirroring the algorithm itself, the following description is given in four parts: (i) an entry-wise guarantee for $\hat{M}$; (ii) a moment matching estimator for $\hat{\theta}$; (iii) a confidence interval for f^*_{ij} ; (iv) an analysis of the optimization problem $\mathcal{P}^{\mathrm{EW}}$.

Step 1: Entry-wise guarantee for $\hat{M}$. Our algorithm is initiated with a de-noising of X_Ω. To ease notation, let $\theta = (p_{\mathrm{A}}, \alpha)$ and $\theta^* = (p^*_{\mathrm{A}}, \alpha^*) \in \Theta$ and denote $e(\theta) := p_{\mathrm{A}} g(\alpha) + (1 - p_{\mathrm{A}})$. This latter function is chosen so that, as follows from a quick calculation, $\mathbb{E}\left(X\right) = e(\theta^*)M^*$. While the SVD-based de-noising algorithm used here is standard, the key result that drives rest of the algorithm and analysis is the following new *entry-wise* error bound, which may be of independent interest:

Theorem 2. *Let $\hat{M} = \frac{nm}{|\Omega|}\mathrm{SVD}(X_\Omega)_r$. With probability $1 - \frac{1}{nm}$,*

$$\left\| \hat{M} - e(\theta^*)M^* \right\|_{\max} \leq \frac{C(\kappa^4 \mu r)L}{p_{\mathrm{O}}} \frac{\log m\sqrt{m}}{n}.$$

In contrast to previous aggregate error bounds for sub-exponential noise, such as the recent Frobenius norm bound in [16], Theorem 2 implies that $\hat{M}$ is close to its expectation *at each entry*. This enables us in the next steps to infer both the parameters θ^* and the posterior probabilities of anomalies at each entry. The proof of Theorem 2, which can be found in the Appendix, is based on recently-developed techniques for entry-wise analysis of random matrices [13]. Originally used for sub-gaussian noise, we apply those techniques to sub-exponential distributions using Bernstein-type inequalities and a generalization of a recent matrix completion result for Poisson observations [16].

Step 2: Moment matching estimator. Step 1 yields an (entry-wise) accurate estimator $\hat{M}$ of M^*, but only up to some linear scaling that depends on the unknown anomaly model parameters θ^*. Now in Step 2, we are able to use $\hat{M}$ to estimate that unknown scaling $e(\theta^*)$, along with θ^* itself, via a generalized moment of the cumulative distribution function at sufficiently many values for identifiability. In particular, for any $t \in \mathbb{Z}^+$, let $g_t(\theta, M)$ be the proportion of entries of X_Ω expected to be at most t:

$$\begin{aligned} g_t(\theta, M) &:= \mathbb{E}\left(|X_{ij} \leq t, (i,j) \in \Omega|\right) / \mathbb{E}\left(|\Omega|\right) \\ &= \frac{1}{nm} \sum_{(i,j)\in[n]\times[m]} \left(p_{\mathrm{A}} \mathbb{P}_{\mathrm{Anom}}\left(X_{ij} \leq t | \alpha, M_{ij}\right) + (1 - p_{\mathrm{A}})\mathbb{P}_{\mathrm{Poisson}}\left(X_{ij} \leq t | M_{ij}\right)\right). \end{aligned}$$

Given that $M^* \approx \hat{M}/e(\theta^*)$, we choose $\hat{\theta}$ to be the minimizer of the following function which seeks to match a set of T empirical moments to their expectations as closely as possible (in ℓ^2 distance),

$$\hat{\theta} := \arg\min_{\theta\in\Theta} \sum_{t=0}^{T-1} \left(g_t(\theta, \hat{M}/e(\theta)) - |X_{ij} = t, (i,j) \in \Omega|/|\Omega| \right)^2, \tag{4}$$

where T is a large enough constant for identifiability (usually $T = d + 1$ for $\theta \in \mathbb{R}^{d+1}$).

At this point, we can formally state the additional regularity conditions that we require. Let $F = (F_0, F_1, \ldots, F_{T-1}) : \Theta \to \mathbb{R}^T$ be defined as $F_t(\theta) = g_t(\theta, M^* e(\theta^*)/e(\theta))$. Let $\delta' = \frac{(\kappa^4 \mu r)L}{p_{\mathrm{O}}} \frac{\log m\sqrt{m}}{n}$ be the entry-wise bound of $\left\| \hat{M} - e(\theta^*)M^* \right\|_{\max}$.

(RC) Regularity Conditions on $F(\theta)$:

- $F : \Theta \to \mathbb{R}^T$ is continuously differentiable and injective.
- For any $\theta \in \Theta$, $\| J_F(\theta) - J_F(\theta^*) \|_2 \leq \frac{C}{\delta' \log(n)} \|\theta - \theta^*\|$ where J is the Jacobian matrix.
- $\left\| J_F(\theta^*)^{-1} \right\|_2 \leq C$.
- $B_{\delta' \log(n)}(\theta^*) \subset \Theta$ where $B_r(\theta^*) = \{\theta : \|\theta^* - \theta\| \leq r\}$.

These conditions are among the typical set of conditions for methods involving generalized moments. Assuming these conditions hold, the following Lemma establishes that our moment matching estimator is able to accurately estimate θ^*

Lemma 1. *Assuming the above regularity conditions on $F(\theta)$, with probability $1 - \frac{1}{nm}$,*

$$\left\| \hat{\theta} - \theta^* \right\| \le C \frac{(K+L)(\kappa^4 \mu r) L}{p_{\mathrm{O}}} \frac{\log m \sqrt{m}}{n}.$$

Proof Sketch. First, note that both $|X_{ij} \le t, (i,j) \in \Omega|$ and $|\Omega|$ concentrate rapidly to their respective expectations since both are sums of independent Bernoulli variables. Also, $g_t(\theta, M)$ is Lipschitz with respect to M, due to the assumed Lipschitz continuity of $\mathbb{P}_{\mathrm{Anom}}$ and $\mathbb{P}_{\mathrm{Poisson}}$. Hence $g_t(\theta^*, \hat{M}/e(\theta^*)) \approx g_t(\theta^*, M^*) \approx |X_{ij} \le t, (i,j) \in \Omega|/|\Omega|$. Since $\hat{\theta}$ is the optimizer of Eq. (4), we have $g_t(\hat{\theta}, \hat{M}/e(\hat{\theta})) \approx |X_{ij} \le t, (i,j) \in \Omega|/|\Omega|$. Lipschitz continuity of $g_t(\theta, M)$ on M then gives us that $g_t(\hat{\theta}, M^* e(\theta^*)/e(\hat{\theta})) \approx g_t(\hat{\theta}, \hat{M}/e(\hat{\theta})) \approx g_t(\theta^*, M^*)$. This implies $F(\theta^*) \approx F(\hat{\theta})$, which along with our regularity conditions, finally implies $\theta_1 \approx \theta_2$. See Appendix for further details. $\square$

Before proceeding, a possible question here is why a more 'natural' estimator such as the MLE was not used. The reason is that our estimator needs to be, in a sense, robust to model misspecification as a result of using $\hat{M}$ as a proxy for $e(\theta^*)M^*$. The MLE does not have this property here, loosely due to the unboundedness of the KL-divergence between two Poisson distributions. On the other hand, the estimator we have proposed indeed provides the desired robustness.

Step 3: Confidence interval. Next, we estimate a confidence interval $[f_{ij}^{\mathrm{L}}, f_{ij}^{\mathrm{R}}]$ for each conditional anomaly probability $f_{ij}^*, (i,j) \in \Omega$ using what effectively amounts to a plug-in estimator along with the high-probability guarantee of Lemma 1. Let $\hat{x}_{ij} := [\hat{p}_{\mathrm{A}} \mathbb{P}_{\mathrm{Anom}}(X_{ij} | \hat{\alpha}, \hat{M}_{ij}/e(\hat{\theta}))]$, $\hat{y}_{ij} := [(1-\hat{p}_{\mathrm{A}}) \mathbb{P}_{\mathrm{Poisson}}(X_{ij} | \hat{M}_{ij}/e(\hat{\theta}))]$ where $[x]$ denotes x 'truncated' to its nearest value in $[0,1]$, i.e. $[x] = \max(\min(x,1),0)$.

By Lemma 1, $\|\hat{\theta} - \theta^*\| \lesssim \delta/(K+L)$, where $\delta := \frac{(K+L)^2(\kappa^4 \mu r)L}{p_{\mathrm{O}}} \frac{\log m \sqrt{m}}{n}$. Given this, along with Lipschitz continuity of the density function, we could expect that $\hat{x}_{ij}, \hat{y}_{ij}$ are sufficiently 'close' to $x_{ij} = p_{\mathrm{A}}^* \mathbb{P}_{\mathrm{Anom}}\left(X_{ij} | \alpha^*, M_{ij}^*\right), y_{ij} = (1-p_{\mathrm{A}}^*) \mathbb{P}_{\mathrm{Poisson}}\left(X_{ij} | M_{ij}^*\right)$, so that they might yield an accurate confidence interval of $f_{ij}^* = y_{ij}^*/(x_{ij}^* + y_{ij}^*)$. That is the following result:

Lemma 2. *There exists a (known) constant C_1 such that, if*

$$f_{ij}^{\mathrm{L}} := \left[\frac{\hat{y}_{ij} - C_1 \delta}{\hat{x}_{ij} + \hat{y}_{ij}} \right] \quad \text{and} \quad f_{ij}^{\mathrm{R}} := \left[\frac{\hat{y}_{ij} + C_1 \delta}{\hat{x}_{ij} + \hat{y}_{ij}} \right], \tag{5}$$

then with probability $1 - \frac{1}{nm}$, we have $f_{ij}^{\mathrm{L}} \le f_{ij}^* \le f_{ij}^{\mathrm{L}} + \epsilon_{ij}$, and $f_{ij}^{\mathrm{R}} - \epsilon_{ij} \le f_{ij}^* \le f_{ij}^{\mathrm{R}}$, where $\epsilon_{ij} = \min(4C_1\delta/(x_{ij} + y_{ij}), 1)$.

Steps 4-5: The optimization problem $\mathcal{P}^{\mathrm{EW}}$. The final two steps involve solving $\mathcal{P}^{\mathrm{EW}}$. To motivate its particular form, consider the 'ideal' anomaly detection algorithm if the f_{ij}^*'s were known. Intuitively, one should claim anomalies at entries with the smallest values of f_{ij}^*. This leads to the following idealized algorithm, which we will call $\pi^*(\gamma)$:

1. Let $\{t_{ij}^*\}$ be an optimal solution to the following optimization problem.

$$\begin{aligned} \mathcal{P}^* : \max_{\{0 \le t_{ij} \le 1, (i,j) \in \Omega\}} & \sum_{(i,j) \in \Omega} t_{ij} \\ \text{subject to} & \sum_{(i,j) \in \Omega} t_{ij} f_{ij}^* \le \gamma \sum_{(i,j) \in \Omega} f_{ij}^* \end{aligned}$$

2. For every $(i,j) \in \Omega$, generate $A_{ij} \sim \mathrm{Ber}(t_{ij}^*)$ independently.

For any algorithm π and any observation X_Ω, let $t_{ij}^\pi(X_\Omega) := \mathbb{P}\left(A_{ij}^\pi(X) = 1\right)$. If $\mathrm{FPR}_\pi(X_\Omega) \le \gamma$, then $\{t_{ij}^\pi(X_\Omega)\}$ is a feasible solution of $\mathcal{P}^*$ by Eq. (3). Furthermore, the objective value of $\mathcal{P}^*$ at the point $\{t_{ij}^\pi(X_\Omega)\}$ is positive correlated with $\mathrm{TPR}_\pi(X_\Omega)$ by Eq. (2). One can verify that this yields the following claim:

Claim 3.1. *If $\mathrm{FPR}_\pi(X_\Omega) \le \gamma$, then $\mathrm{TPR}_\pi(X_\Omega) \le \mathrm{TPR}_{\pi^*(\gamma)}(X_\Omega)$.*

Now notice that $\mathcal{P}^{\mathrm{EW}}$ is obtained by replacing f^*_{ij} with the confidence interval estimators f^{L}_{ij} and f^{R}_{ij} defined in the previous step. Intuitively, we could expect that $\mathcal{P}^{\mathrm{EW}} \approx \mathcal{P}^*$, and therefore the algorithm π^{EW} should achieve the desired performance. In fact, $\mathrm{FPR}_{\pi^{\mathrm{EW}}(\gamma)}(X) \leq \gamma$ holds immediately because $f^{\mathrm{L}}_{ij} \leq f^*_{ij} \leq f^{\mathrm{R}}_{ij}$ and so $\{t^{\mathrm{EW}}_{ij}\}$ is a feasible solution of $\mathcal{P}^*$. To show the desired performance guarantee for $\mathrm{TPR}_{\pi^{\mathrm{EW}}}(X)$, we first prove the following Lemma:

Lemma 3. *With probability* $1 - \frac{1}{nm}$, $\sum_{(i,j)\in\Omega}(|f^L_{ij} - f^*_{ij}| + |f^R_{ij} - f^*_{ij}|) \leq CL\log(1/\delta)\delta p_{\mathrm{O}} nm.$

Let $\{t'_{ij}\}$ be the optimal solution of $\pi^*(\gamma')$ where $\frac{\sum_{(i,j)\in\Omega} t'_{ij}}{\sum_{(i,j)\in\Omega} t^*_{ij}} = \eta < 1$. The key idea is to find some η such that $\{t'_{ij}\}$ is a feasible solution of $\mathcal{P}^{\mathrm{EW}}$, while maintaining good performance compared to $\pi^*(\gamma)$. Indeed, a sufficiently large η can be achieved by Lemma 3. In particular, we have:

Lemma 4. *Let* $\eta = CL\delta\log(1/\delta)$. *Then* $\{t'_{ij}\}$ *is a feasible solution of* $\mathcal{P}^{\mathrm{EW}}$. *Furthermore,* $\frac{\sum_{(i,j)\in\Omega} t^*_{ij} - \sum_{(i,j)\in\Omega} t'_{ij}}{\sum_{(i,j)\in\Omega}(1-f^*_{ij})} \leq C_1 \frac{L\delta\log(1/\delta)}{\gamma p^*_{\mathrm{A}}}$ *for a constant* C_1.

Finally, some algebra gives us that $\mathrm{TPR}_{\pi^*(\gamma)} - \mathrm{TPR}_{\pi^{\mathrm{EW}}(\gamma)} \leq \frac{\sum_{(i,j)\in\Omega} t^*_{ij} - \sum_{(i,j)\in\Omega} t^{\mathrm{EW}}_{ij}}{\sum_{(i,j)\in\Omega}(1-f^*_{ij})}$. Applying Lemma 4, this completes the proof (sketch) of Theorem 1, since $\sum_{(i,j)\in\Omega} t^{\mathrm{EW}}_{ij} \geq \sum_{(i,j)\in\Omega} t'_{ij}$ because $\{t^{\mathrm{EW}}_{ij}\}$ is the optimal solution of $\mathcal{P}^{\mathrm{EW}}$.

3.2 Minimax Lower Bound

In this final subsection, we provide a minimax lower bound on the regret of TPR, which confirms that Theorem 1 is optimal up to logarithmic terms. To do this, we construct the following simple model: let $p_{\mathrm{O}} = 1$, and when an anomaly occurs, assume $X_{ij} = 0$. We refer to this in notational form as $X \sim \mathrm{Q}(p^*_{\mathrm{A}}, M^*)$. Now we construct a set of matrices $\mathcal{M}_n = \{M^b \in \mathbb{R}^{n\times n}, b \in \{0,1\}^{n/2}\}$ as follows. For the i-th and $(i+1)$-th rows, set $M_{ij} = 1$ and $M_{i+1j} = 1 - \frac{C}{\sqrt{n}}$ if $b_{i/2} = 0$; otherwise set $M_{ij} = 1 - \frac{C}{\sqrt{n}}$ and $M_{i+1j} = 1$. Here C is some constant. One can verify that $K = L = \mu = r = \kappa = O(1)$ for $X \sim \mathrm{Q}(p^*_{\mathrm{A}}, M^b)$ where $M^b \in \mathcal{M}_n$.

In the following proposition, we show that even for this simple anomaly model, one cannot expect regret on TPR better than $O(1/\sqrt{n})$. To allow for comparison to Theorem 1, let Π_γ denote the set of all policies such that

$$\mathbb{P}_{X\sim\mathrm{Q}(p^*_{\mathrm{A}},M)}\left(\mathrm{FPR}_\pi(X) \leq \gamma\right) \geq 1 - C/n^2 \quad \text{for all } M \in \mathcal{M}_n.$$

Proposition 1. *Let* $\gamma = \frac{1}{2e}$ *and* $p^*_{\mathrm{A}} = \frac{1}{2}$. *For any algorithm* $\pi \in \Pi_\gamma$, *there exists* $M' \in \mathcal{M}_n$ *such that*

$$\mathbb{E}_{X\sim\mathrm{Q}(p^*_{\mathrm{A}},M')}\left(\mathrm{TPR}_{\pi^*(\gamma)}(X) - \mathrm{TPR}_\pi(X)\right) \geq C/\sqrt{n}.$$

4 Experiments

In studying the empirical performance of the EW, we first consider a synthetic setting where we examine the impact of natural problem parameters on performance. We measure the AUC achieved by EW, and how it compares to an AUC upper bound as well as the AUC of Stable PCP (a state-of-the-art approach). We then study performance on real world data from a large CPG research partner.

Synthetic Data: We consider generating an ensemble of M^* matrices: Let $n = m = 100$. For a given choice of r and entry-wise mean $\bar{M}^*$, we set $M^* = kUV^T$. $U, V \in \mathbb{R}^{n\times r}$ are random with independent $\mathrm{Gamma}(1,2)$ entries and k is picked so that $\bar{M}^* = \frac{1}{nm}\sum_{ij} M^*_{ij}$. If (i,j) is observed, then X_{ij} is Poisson with mean M_{ij} with probability $1 - p^*_{\mathrm{A}}$; otherwise, it is Poisson with mean $a_{ij}M_{ij}$ where a_{ij} is exponentially distributed with mean α^*. We consider an ensemble of 1000 problems obtained by uniformly drawing $r \in [1,10], \bar{M}^* \in [1,10], p_{\mathrm{O}} \in [0.5,1], p^*_{\mathrm{A}} \in [0,0.3]$ and $\alpha^* \in [0,1]$. We consider an implementation of the EW algorithm where the matrix completion step used the soft impute algorithm [27] and the anomaly model estimation used MLE. Stable PCP solves Eq. (1). In both cases, we tuned Lagrange multipliers corresponding to rank using knowledge of

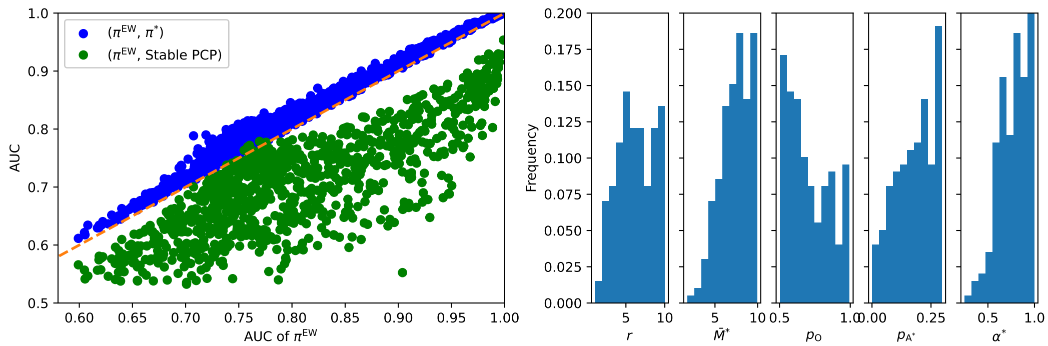

Figure 1: Synthetic data. Left scatter shows AUC of ideal algorithm vs that of EW (blue points, above 45-degree line); and AUC of Stable PCP vs EW (green, mostly below 45 degree line). Histograms shows problem characteristics where EW performs worst relative to ideal (20th percentile).

the true rank. For convex optimization, we generated an ROC curve for each problem instance by varying the Lagrange multiplier penalizing $\|A\|_1$; for EW we do this by simply varying γ.

Left of the Figure 1 shows that EW consistently achieves an AUC close to that of a super-optimal algorithm ('ideal', that knows M^* and the anomaly model) while Stable PCP is substantially worse than EW. Right of the Figure 1 shows that the problem instances where the AUC of EW was furthest away from the ideal AUC show largely intuitive characteristics: higher α^* (so anomalies look similar to non-anomalous entires), lower p_O, higher p_A^* and higher r (so that M^* is harder to estimate). The behavior with respect to $\bar{M}^*$ is surprising but was consistently observed across other ensembles.

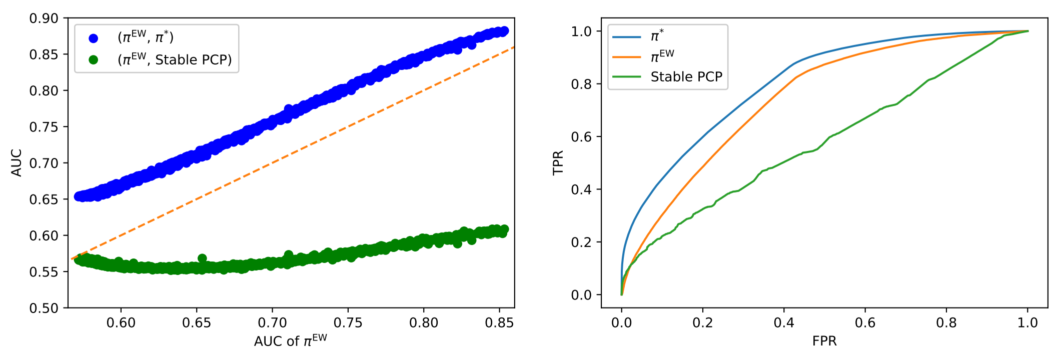

Figure 2: Real data. The left display considers an ensemble similar to the synthetic data. Right display corresponds to a representive setting of $p_A^* = 0.04$ and $\alpha^* = 0.2$.

Real Data: This consists of sales of $m = 290$ products across $n = 2481$ stores with $p_O \sim 0.14$. M^* is obtained by denoising this data with $r \sim 30$ (estimated via cross-validation). Average observed sales per product-store was $\bar{M}^* = 2.64$; so the variance of non-anomalous entries is relatively large. We generate X as before, introducing anomalies by deliberately perturbing a fraction p_A^* of entries and thinning the resulted sales at rate α^*. We generate an ensemble of 1000 such perturbed matrices.

Figure 2 considers the ensemble of perturbed matrices; we see similar relative merits as in the synthetic experiments: EW achieves an AUC close to that of an algorithm that knows M^* and α^* whereas Stable PCP is consistently worse than EW. Right of the Figure 2 shows an AUC curve for a representative setting of $p_A^* = 0.04$ and $\alpha^* = 0.2$ where we see the absolute performance: the AUC for the ideal algorithm was ~ 0.806 whereas the AUC for EW was close at 0.747 – this suggests that EW is quite viable in this domain. Stable PCP is substantially worse with an AUC of 0.58.

Broader Impact

The primary motivation for this work is the phantom inventory problem for retailers. Given the dramatic cost of this problem, we anticipate algorithmic approaches to addressing the problem are of potential commercial value.

References

[1] Zihan Zhou, Xiaodong Li, John Wright, Emmanuel Candes, and Yi Ma. Stable principal component pursuit. In *2010 IEEE international symposium on information theory*, pages 1518–1522. IEEE, 2010.

[2] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[3] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[4] Zhouchen Lin, Arvind Ganesh, John Wright, Leqin Wu, Minming Chen, and Yi Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *Coordinated Science Laboratory Report no. UILU-ENG-09-2214, DC-246*, 2009.

[5] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[6] Xiaoming Yuan and Junfeng Yang. Sparse and low-rank matrix decomposition via alternating direction methods. *preprint*, 12(2), 2009.

[7] Necdet Serhat Aybat. Algorithms for stable pca. *Handbook of Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image and Video Processing*, pages 2–1, 2016.

[8] Shiqian Ma and Necdet Serhat Aybat. Efficient optimization algorithms for robust principal component analysis and its variants. *Proceedings of the IEEE*, 106(8):1411–1426, 2018.

[9] Huishuai Zhang, Yi Zhou, and Yingbin Liang. Analysis of robust pca via local incoherence. In *Advances in Neural Information Processing Systems*, pages 1819–1827, 2015.

[10] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[11] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust pca via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

[12] Teng Zhang and Yi Yang. Robust pca by manifold optimization. *The Journal of Machine Learning Research*, 19(1):3101–3139, 2018.

[13] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

[14] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *arXiv preprint arXiv:1906.04159*, 2019.

[15] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, pages 1–182, 2019.

[16] Andrew D McRae and Mark A Davenport. Low-rank matrix completion and denoising under poisson noise. *arXiv preprint arXiv:1907.05325*, 2019.

[17] Ananth Raman, Nicole DeHoratius, and Zeynep Ton. Execution: The missing link in retail operations. *California Management Review*, 43(3):136–152, 2001.

[18] Nicole DeHoratius and Ananth Raman. Inventory record inaccuracy: An empirical analysis. *Management science*, 54(4):627–641, 2008.

[19] Ti-Jun Fan, Xiang-Yun Chang, Chun-Hua Gu, Jian-Jun Yi, and Sheng Deng. Benefits of rfid technology for reducing inventory shrinkage. *International Journal of Production Economics*, 147:659–665, 2014.

[20] Fuqiang Wang, Xiaoping Fang, Xiaohong Chen, and Xihua Li. Impact of inventory inaccuracies on products with inventory-dependent demand. *International Journal of Production Economics*, 177:118–130, 2016.

[21] Heather Nachtmann, Matthew A Waller, and David W Rieske. The impact of point-of-sale data inaccuracy and inventory record data errors. *Journal of Business Logistics*, 31(1):149–158, 2010.

[22] Li Chen and Adam J Mersereau. Analytics for operational visibility in the retail store: The cases of censored demand and inventory record inaccuracy. In *Retail supply chain management*, pages 79–112. Springer, 2015.

[23] A Gürhan Kök and Kevin H Shang. Inspection and replenishment policies for systems with inventory record inaccuracy. *Manufacturing & service operations management*, 9(2):185–205, 2007.

[24] Nicole DeHoratius, Adam J Mersereau, and Linus Schrage. Retail inventory management when records are inaccurate. *Manufacturing & Service Operations Management*, 10(2):257–277, 2008.

[25] SA Conrad. Sales data and the estimation of demand. *Journal of the Operational Research Society*, 27(1):123–127, 1976.

[26] Jim Shi, Michael N Katehakis, Benjamin Melamed, and Yusen Xia. Production-inventory systems with lost sales and compound poisson demands. *Operations Research*, 62(5):1048–1063, 2014.

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