Near-Optimal Entrywise Anomaly Detection for Low-Rank Matrices with Sub-Exponential Noise

Vivek F. Farias [1] **Andrew A. Li** [2] **Tianyi Peng** [3]

Abstract

We study the problem of identifying anomalies in a low-rank matrix observed with sub-exponential noise, motivated by applications in retail and inventory management. State of the art approaches to anomaly detection in low-rank matrices apparently fall short, since they require that non-anomalous entries be observed with vanishingly small noise (which is not the case in our problem, and indeed in many applications). So motivated, we propose a conceptually simple entrywise approach to anomaly detection in low-rank matrices. Our approach accommodates a general class of probabilistic anomaly models. We extend recent work on entrywise 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. Viewing the anomaly detection as a classification task, to the best of our knowledge, we are the first to achieve the min-max optimal detection rate (up to log factors). 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 $X = M^* + E + A$, where E is a *noise* matrix with independent, mean-zero entries, and where A is a sparse matrix of *anomalies*. We observe only X, and only on some subset of matrix entries Ω. The anomaly detection problem concerns identifying the support of A simply from this observation.

One of the most popular approaches to solving this problem is the following convex optimization formulation, referred to as 'stable principal component pursuit' (stable PCP) (Zhou et al., 2010),

$$\min_{\hat{M},\hat{A}} \|\hat{M}\|_* + \lambda_2\|\hat{A}\|_1 + \lambda_1\|P_\Omega(X - \hat{M} - \hat{A})\|_{\mathrm{F}}^2, \quad (1)$$

where λ_1 and λ_2 are regularization parameters. The three matrix norms in the objective are meant to promote, from left to right, low-rankedness in $\hat{M}$, sparsity in $\hat{A}$, and fit to X on the observed entries Ω. Upon solving problem (1), $\hat{A}$ can be used to estimate the support of A. Now in the absence of anomalies, this optimization problem (after removing the $\hat{A}$ terms) is in essence optimal under a variety of assumptions on the distributions of E and Ω. In contrast, the available results for anomaly detection are weaker. Perhaps most limiting, results that guarantee recovery of M^* require that the 'average' noise $\|E\|_{\mathrm{F}}/n$ vanishes, where n is 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. X is typically noisy: In the practical problem that motivates this work, $M^* + E$ can be viewed as a matrix of centered Poisson entries with mean M^*. Clearly then, $\mathbb{E}\|E\|_{\mathrm{F}}$ 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 X is noisy (such as in our motivating application), the optimization approach above can perform quite poorly.

1.1. Overview of Main Contributions

Against the above backdrop, we make the following contributions to the problem of anomaly detection in matrices:

1. **An optimal algorithm:** We develop a new algorithm for low-rank matrices with sub-exponential noise, and

[1]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [2]Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213, USA [3]Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Correspondence to: Tianyi Peng <tianyi@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

prove that our approach, for the first time, achieves the min-max optimal anomaly detection rate (up to logarithmic terms) under a broad class of probabilistic anomaly models (Theorem 1 and Proposition 1).

2. **Entrywise guarantee for sub-exponential matrices:** As part of our approach, we prove a new recovery guarantee of independent interest (Theorem 2) for the matrix completion problem. This result is unique in applying to sub-exponential (vs. the usual sub-gaussian) noise, and bounding the *entrywise* (vs. an aggregate) error.

3. **Applications:** Our work is motivated by a crucial inventory management problem ('phantom inventory') 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 time. On large-scale data (thousands of stores, thousands of products) we find that our approach achieves 13% and 19% higher accuracy (measured via the usual area under ROC curve) than convex optimization approaches on synthetic and real data, respectively.

Why sub-exponential noise: Sub-exponentiality relaxes the typical sub-gaussianity assumption, and includes a broader class of relevant distributions such as the Poisson, exponential, and Chi-square distributions. In particular, sub-exponentiality occurs frequently in applications, e.g. to model sales and demand. The technical difficulty mainly arises from deriving the proper concentration bounds for high-dimensional matrices.

1.2. 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 existing work has focused on a formulation called *robust principal component analysis (robust PCA)* (Candès et al., 2011; Chandrasekaran et al., 2011). Most relevant to our problem are approaches for *noisy* robust PCA (Zhou et al., 2010; Agarwal et al., 2012; Wong & Lee, 2017; Klopp et al., 2017; Chen et al., 2020b). See Table 1 for a summary of existing results. Note that any hope of identifying the anomalies A would require, at the very least, that $\|\hat{M} - M^*\|_{\mathrm{F}} = o(n)$. Thus, with respect to the noisy problem we are studying, in which $\|E\| = \Omega(n)$, existing results are insufficient. In contrast, our algorithm not only improves upon the recovery of M^* to sufficiently allow for recovery of A, it also provides an additional guarantee on entrywise recovery: $\|\hat{M} - M^*\|_{\max}$. All our guarantees are min-max optimal, and beyond the recovery of M^*, to the best of our knowledge, we are also the first paper to analyze the anomaly detection *rate* as a formal classification problem. Finally, concurrent with this paper, (Chen et al., 2020b) provide a more-refined analysis of Eq. (1) and achieve a recovery guarantee for M^* similar to ours. However, they require a zero-mean assumption on the anomalies, i.e. $\mathbb{E}(A) = 0$, which is not representative of many applications, such as the ones motivating this work.[1]

	$\left\|\hat{M} - M^*\right\|_{\mathrm{F}}$	$\left\|\hat{M} - M^*\right\|_{\max}$
(Zhou et al., 2010)	$n\,\|E\|_{\mathrm{F}}$	–
(Wong & Lee, 2017)	$\sqrt{n}\,\|E\|_{\mathrm{F}}$	–
(Klopp et al., 2017)	$\sqrt{\log n}\,\|E\|_{\mathrm{F}}$	–
This paper	$\frac{\sqrt{\log n}}{\sqrt{n}}\,\|E\|_{\mathrm{F}}$	$\frac{\sqrt{\log n}}{n\sqrt{n}}\,\|E\|_{\mathrm{F}}$

Table 1: Comparison of our results with existing work under proper hyper-parameters. The reported quantities are the scalings of upper bounds on the error of $\|\hat{M} - M^*\|$, for two matrix norms, with respect to matrix size n and noise E.

The second body of work concerns statistical inference in matrix completion (Abbe et al., 2017; Chen et al., 2019; Ma et al., 2019; Chen et al., 2020a). Our contribution here is an entrywise error bound for matrix completion under sub-exponential noise. This result substantially improves upon previous results for sub-exponential matrices, all of which bound an aggregate error measure (Lafond, 2015; Sambasivan & Haupt, 2018; Cao & Xie, 2015; McRae & Davenport, 2019). Our analysis builds on a recent framework introduced in (Abbe et al., 2017) for sub-gaussian noise, and requires both a considerably more fine-tuned computation, and drawing from a recent result for Poisson matrix completion from (McRae & Davenport, 2019).

Finally, with respect to our motivating application: the phantom inventory problem is well-studied in the field of Operations Management (Raman et al., 2001; DeHoratius & Raman, 2008; Nachtmann et al., 2010; Fan et al., 2014; Chen & Mersereau, 2015; Wang et al., 2016). Existing algorithmic solutions (Kök & Shang, 2007; DeHoratius et al., 2008) have focused on adapting inventory management policies to this issue. Algorithmic *detection*, particularly in a form that combines observations across products and stores, is unstudied and is the motivation for this work.

Notation: The sub-exponential norm of X is defined as $\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}(\exp(|X|/t)) \leq 2\}$. For

[1] As pointed out in (Chen et al., 2020b), this assumption is likely a fundamental limitation of Eq. (1).

$A \in \mathbb{R}^{n\times m}$, we write $\sum_{(i,j)\in[n]\times[m]} A_{ij}$ as $\sum_{ij} A_{ij}$ when no ambiguity exists. We require various matrix norms: $\|A\|_{2,\infty}^2 = \max_i \sum_j A_{ij}^2$, $\|A\|_{\max} = \max_{ij} |A_{ij}|$, $\|A\|_{\mathrm{F}}^2 = \sum_{ij} A_{ij}^2$, $\|A\|_0 = \sum_{ij} \mathbb{1}\{A_{ij} \neq 0\}$. The spectral norm of A is denoted $\|A\|_2$. The letter C (or c) represents a sufficiently large (or small) universal (i.e. independent of problem parameters) constant that may change between equations.

2. Model

The following is the core problem we will study. There exists an (unobserved) 'rate' matrix $M^* \in \mathbb{R}_+^{n\times m}$ ($n \leq m$ 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, a *random* matrix X is generated with independent entries distributed according to[2]

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

$\mathrm{Anom}(\cdot,\cdot)$ is some non-negative, integer-valued random variable and $\alpha^* \in \Gamma \subset \mathbb{R}^d$ 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 $p_{\mathcal{O}}$. In addition, we assume that B is a $\mathrm{Bernoulli}(p_{\mathrm{A}}^*)$ 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 both M^* and the anomaly distribution.

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 (Conrad, 1976; Shi et al., 2014). Our results can be easily generalized from Poisson distribution to general sub-exponential distribution. 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 $\mathrm{Anom}(\alpha^*, \lambda)$ is perhaps best viewed as a censored $\mathrm{Poisson}(\lambda)$ random variable. 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 $\sigma_1^* \geq \sigma_2^* \geq \ldots \geq \sigma_r^*$ ($\kappa = \sigma_1^*/\sigma_r^*$); 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): $\|M^*\|_{\max} + 1 \leq L$.
- (Incoherence): $\|U\|_{2,\infty} + \|V\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n+m}}$.
- (Sparsity): $\sqrt{p_{\mathcal{O}}} \geq C_1 \frac{\log^{1.5}(m)\mu r L \kappa^2}{\|M^*\|_{\max}\sqrt{m}}$ for some known constant C_1.

Our guarantees will be parameterized by μ, L, r and κ. These assumptions and parameters for M^* are similar to those in the existing matrix completion literature (Abbe et al., 2017; Ma et al., 2019).

Assumptions on $\mathrm{Anom}(\cdot,\cdot)$: We make the following assumptions:

- (Sub-exponential): $\mathrm{Anom}(\alpha^*, M^*_{ij})$ is sub-exponential: $\left\|\mathrm{Anom}(\alpha^*, M^*_{ij})\right\|_{\psi_1} \leq L$.
- (Lipschitz): For any $M \in \mathbb{R}_+, \alpha \in \Gamma$ and all $k \in \mathbb{N}$, $\mathbb{P}(\mathrm{Anom}(\alpha, M) = k)$ is K-Lipschitz in (α, M).
- (Mean Decomposition): For any $M \in \mathbb{R}_+, \alpha \in \Gamma$, we have $\mathbb{E}(\mathrm{Anom}(\alpha, M)) = g(\alpha)M$ for some $g : \mathbb{R}^d \to \mathbb{R}$ where $g(\alpha)$ is K-Lipschitz in α.

We pause to discuss these assumptions on anomalies. To begin, we assume a probabilistic anomaly model characterized by finite unknown parameters α. This generally applies to many applications. The probabilistic and Lipschitz properties enable the identification of α and then the measure of the anomaly detection rate. We also assume a mean-decomposition condition: $\mathbb{E}(A) \propto M^*$. Note that this is less restrictive than the zero-mean assumption $\mathbb{E}(A) = 0$ sometimes found in the literature. It is also well justified from the known mechanism for anomalies in the phantom inventory problem such as $\mathrm{Anom}(\alpha, M^*_{ij}) = \mathrm{Poisson}(\alpha M^*_{ij})$. (DeHoratius et al., 2008).

In contrast to this probabilistic model, one could consider an adversarial anomaly model. Note that the adversarial model that allows for arbitrary anomalies requires in essence *exact* observations of M^* for non-anomalous observations (Candès et al., 2011), which is not consistent with our highly noisy setup.

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, $\mathrm{TPR}_\pi(X_\Omega)$ 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, $\mathrm{FPR}_\pi(X_\Omega)$. More formally, let f_{ij}^* be the conditional probability that the (i,j)-th entry is not anomalous, given X, i.e.

$$f_{ij}^* := \mathbb{P}(B_{ij} = 0 \mid X).$$

[2] We focus here on a model with non-negative integer-valued X that fits our application. The results can easily be extended to more general sub-exponential noise, as we discuss in Section 4.2.

Then, some algebraic manipulation establishes that

$$\mathrm{TPR}_\pi(X_\Omega) = \frac{\sum_{(i,j)\in\Omega} \mathbb{P}\left(A^\pi_{ij}(X_\Omega)=1\right)(1-f^*_{ij})}{\sum_{(i,j)\in\Omega}(1-f^*_{ij})}$$
$$\mathrm{FPR}_\pi(X_\Omega) = \frac{\sum_{(i,j)\in\Omega} \mathbb{P}\left(A^\pi_{ij}(X_\Omega)=1\right) f^*_{ij}}{\sum_{(i,j)\in\Omega} f^*_{ij}}.$$

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 (clairvoyant) optimal estimator. We will show that in large matrices this gap grows negligibly small at a min-max optimal rate.

3. Algorithm and Results

We are now prepared to state our approach to the anomaly detection problem formulated above. Our algorithm, which we refer to as the *entrywise (EW)* algorithm, leverages an entrywise 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 crux of our algorithm is stated in Algorithm 1 below:

Algorithm 1 Entrywise (EW) Algorithm $\pi^{\mathrm{EW}}(\gamma)$

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

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

2: Estimate $(\hat{p}_{\mathrm{A}}, \hat{\alpha})$ based on a moment matching estimator.

3: Estimate a confidence interval $[f^{\mathrm{L}}_{ij}, f^{\mathrm{R}}_{ij}]$ for f^*_{ij} for $(i,j)\in\Omega$.

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

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

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

Output: A_Ω

Step 2 and Step 3 are based on natural plug-in estimators (see full details specified in Eq. (2) and Eq. (3) respectively). 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) naturally imposed by the estimation of $(\hat{p}_A, \hat{\alpha})$, stated in Section 4.2.

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

$$\mathrm{FPR}_{\pi^{\mathrm{EW}}(\gamma)}(X_\Omega) \le \gamma,$$
$$\mathrm{TPR}_{\pi^{\mathrm{EW}}(\gamma)}(X_\Omega) \ge \mathrm{TPR}_{\pi^*(\gamma)}(X_\Omega) - C\frac{(K+L)^3L^3\kappa^4\mu r}{p^*_{\mathrm{A}}\gamma}\frac{\log^{1.5}(m)}{\sqrt{p_{\mathcal{O}}m}}.$$

To parse this result, consider that in a typical application, we can expect the problem parameters to fall in the following scaling regime: $K, L, \kappa, r, \mu = O(1)$, $p_{\mathcal{O}}, p^*_{\mathrm{A}}, \gamma = \Omega(1)$, and $m/n = \Theta(1)$. For this regime, the regret is $O\left(n^{-1/2}\log^{1.5} n\right)$, which is in fact optimal up to logarithmic factors. To be precise, we fix a particular value of γ for which the following proposition states that, for any n, there exists a family of anomaly models $\mathcal{M}_n$ for which no algorithm can achieve a regret on TPR lower than $O(n^{-1/2})$ across all models within the family (we will explicitly construct this family in the proof in Section 4.4). To allow for direct comparison to Theorem 1, let Π_γ denote the set of all policies π such that

$$\mathbb{P}_{X|M^*}\left(\mathrm{FPR}_\pi(X)\le\gamma\right) \ge 1 - C/n^2 \quad \text{for all } M^*\in\mathcal{M}_n.$$

Proposition 1. *For any algorithm* $\pi\in\Pi_\gamma$*, there exists* $M^*\in\mathcal{M}_n$ *such that*

$$\mathbb{E}_{X|M^*}\left(\mathrm{TPR}_{\pi^*(\gamma)}(X) - \mathrm{TPR}_\pi(X)\right) \ge C/\sqrt{n}.$$

This shows that our algorithm is optimal for the TPR up to logarithmic factors.

Algorithmic Novelty: Before proceeding to the sketches of these results, it is worth discussing the novelty of our algorithm. The vast majority of existing algorithms all seek to decompose X into its three components (M, A, E) by solving a single optimization problem: min $f(M)+g(A)+h(E)$, where f, g, h are carefully chosen penalty functions, e.g., Eq. (1). In contrast to these approaches, our algorithm uses two separate procedures. The first procedure (Step 1) is effectively a de-noising and completion routine which, rather unintuitively, makes no attempt to identify the positions of anomalies, but is able to estimate M with a guarantee on

entrywise accuracy, but only up to an unknown affine scaling. This entrywise guarantee (a fundamentally new result not previously exploited by the aforementioned optimization algorithms) enables the second procedure (Steps 2–5), which leverages the underlying probabilistic structure to estimate the affine scaling and perform *entrywise inference*, yielding the first sharp statements about the optimality of the anomaly detection rate.

4. Algorithm Details and Proof Sketches

In this section, we motivate the steps of Algorithm 1 and sketch the proofs of our main results. Beginning with Theorem 1, and mirroring the algorithm itself, the following description is given in four parts: (i) an entrywise guarantee for $\hat{M}$; (ii) a moment matching estimator for $(\hat{p}_A, \hat{\alpha})$; (iii) a confidence interval for f^*_{ij} ; (iv) an analysis of the optimization problem $\mathcal{P}^{\mathrm{EW}}$.

4.1. Step 1: Entrywise Guarantee for $\hat{M}$

Our algorithm is initiated with a de-noising of X_Ω. To ease notation, let $\theta = (p_{\mathrm{A}}, \alpha)$, $\theta^* = (p^*_{\mathrm{A}}, \alpha^*)$, 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, $\mathbb{E}(X) = e(\theta^*)M^*$. 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 *entrywise* error bound, which may be of independent interest:

Theorem 2. *Let*

$$\hat{M} = \frac{nm}{|\Omega|}\mathrm{SVD}(X_\Omega)_r.$$

Then with probability $1 - O(\frac{1}{nm})$,

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

Our result can be viewed as the first entrywise guarantee result for Poisson matrix completion.[3] The proof sketch is provided in the Appendix. As a comparison, consider the recent results for aggregated error on matrix completion with Poisson noise (McRae & Davenport, 2019). Under the proper hyper-parameters, their results based on SVD provide the following Frobenius norm bound: $\|\hat{M} - M^*\|_{\mathrm{F}} \lesssim n^{1/2}$. In contrast, our entrywise guarantees provide that $\|\hat{M} - M^*\|_{\max} \lesssim n^{-1/2}\log^{1/2} n$. Therefore, our results show that the SVD approach not only provides aggregated error guarantee but also yields a much stronger result: the entrywise error guarantee. Furthermore, the entrywise error is evenly distributed among all entries up to a logarithmic factor.

[3] In fact, the proof also holds valid for sub-exponential noise.

The entrywise guarantee is the key that opens up optimal anomaly detection. In particular, this enables us in the next steps to infer both the parameters θ^* and the posterior probabilities of anomalies at each entry.

4.2. Step 2: Moment Matching Estimator

Step 1 yields an (entrywise) 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 $t \in \mathbb{N}$, let $g_t(\theta, M)$ be the proportion of entries of X_Ω expected to be at most t with the model specified by θ and M:

$$g_t(\theta, M) := \mathbb{E}\left(|X_{ij} \le t, (i,j) \in \Omega|\right) / \mathbb{E}\left(|\Omega|\right).$$

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),

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

where T is a large enough constant for identifiability (usually $T = d + 1$ for $\theta \in \mathbb{R}^{d+1}$) and Θ is chosen such that $\theta^* \in \Theta$ and p_{A} is bounded from 1 by a constant for $\theta = (p_{\mathrm{A}}, \alpha) \in \Theta$.

Let $F = (F_0, F_1, \ldots, F_{T-1}) : \Theta \to \mathbb{R}^T$ be defined as $F_t(\theta) = g_t(\theta, M^* e(\theta^*)/e(\theta))$. We could expect that $F_t(\hat{\theta}) \approx F_t(\theta^*)$ by solving $\hat{\theta}$ from Eq. (2). In fact, we have the following result.

Lemma 1. *With probability* $1 - O(\frac{1}{nm})$,

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

To establish that $\hat{\theta} \approx \theta^*$ from $F(\hat{\theta}) \approx F(\theta^*)$, additional regularity conditions are required. Let $\delta' = \kappa^4 \mu r L \sqrt{\frac{\log m}{p_{\mathcal{O}} m}}$ be the entrywise bound of $\|\hat{M} - e(\theta^*)M^*\|_{\max}$. We now formally state the regularity conditions that we require:

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

- $F : \Theta \to \mathbb{R}^T$ is continuously differentiable and injective.
- Let $\tilde{\delta} = \delta'(K + L)\log m$. We require $B_{\tilde{\delta}}(\theta^*) \subset \Theta$, where $B_r(\theta^*) = \{\theta : \|\theta^* - \theta\| \le r\}$.

- For any $\theta \in B_{\tilde{\delta}}(\theta^*)$, $\|J_F(\theta) - J_F(\theta^*)\|_2 \leq \frac{C}{\tilde{\delta}} \|\theta - \theta^*\|$, where J is the Jacobian matrix.
- $\left\|J_F(\theta^*)^{-1}\right\|_2 \leq C$.

These conditions are among the typical set of conditions for methods involving generalized moments and are well justified in typical applications (Newey & McFadden, 1994; Imbens et al., 1995; Hall, 2005; Hansen, 1982). The following lemma establishes that our moment matching estimator is able to accurately estimate θ^*.

Lemma 2. *Assuming the above regularity conditions* (RC) *on $F(\theta)$, with probability $1 - O(\frac{1}{nm})$,*

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

Extending to general noise models: To extend Algorithm 1 to general sub-exponential noise, all steps hold the same except that the estimator for $\hat{\theta}$ in the Step 2 needs to be changed. For observation X with continuous values, one can use MLE estimator to solve $\hat{\theta} = \arg\max_\theta \mathbb{P}(X|\theta, M^*)$ by plugging in $M^* \approx \hat{M}/e(\theta)$. For integer-value X beyond Poisson noise, we still use moment matching estimator (incorporate negative values in Eq. (2) if needed). In both cases, the results in this paper will still hold with slightly different regularity conditions for the identification of parameters.

4.3. Steps 3–5: Confidence Intervals and the Optimization Problem $\mathcal{P}^{\mathrm{EW}}$:

Next, we use $\hat{M}$ and $\hat{\theta}$ as plug-in estimators to optimize TPR under the FPR constraint. 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. $[x] = \max(\min(x,1),0)$. We then can estimate a confidence interval $[f_{ij}^{\mathrm{L}}, f_{ij}^{\mathrm{R}}]$ for each conditional non-anomaly probability $f_{ij}^* = \mathbb{P}(B_{ij} = 0 \mid X)$ using what effectively amounts to a plug-in estimator based on $\hat{x}_{ij}, \hat{y}_{ij}$. That is the content of the following result:

Lemma 3. *Let*

$$\delta = (K+L)^3 \kappa^4 \mu r L^2 \sqrt{\frac{\log m}{p_{\mathcal{O}} m}}.$$

There exists a (known) constant C_1 such that, if

$$f_{ij}^{\mathrm{L}} := \left[\frac{\hat{y}_{ij} - C_1\delta}{\hat{x}_{ij} + \hat{y}_{ij}}\right] \quad \textit{and} \quad f_{ij}^{\mathrm{R}} := \left[\frac{\hat{y}_{ij} + C_1\delta}{\hat{x}_{ij} + \hat{y}_{ij}}\right], \quad (3)$$

then with probability $1 - O(\frac{1}{nm})$, for every $(i,j) \in \Omega$, we have

$$f_{ij}^{\mathrm{L}} \leq f_{ij}^* \leq f_{ij}^{\mathrm{L}} + \epsilon_{ij} \quad \textit{and} \quad f_{ij}^{\mathrm{R}} - \epsilon_{ij} \leq f_{ij}^* \leq f_{ij}^{\mathrm{R}},$$

where $\epsilon_{ij} = \min(4C_1\delta/(x_{ij}^* + y_{ij}^*), 1)$.

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 identify 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.

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

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

The following claim establishes the optimality of $\pi^*(\gamma)$.

Claim 1. *For any π, γ, and X_Ω, if $\mathrm{FPR}_\pi(X_\Omega) \leq \gamma$, then $\mathrm{TPR}_\pi(X_\Omega) \leq \mathrm{TPR}_{\pi^*(\gamma)}(X_\Omega)$.*

Now notice that $\mathcal{P}^{\mathrm{EW}}$ is obtained from $\mathcal{P}^*$ by replacing f_{ij}^* with the confidence interval estimators f_{ij}^{L} and f_{ij}^{R} 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 $f_{ij}^{\mathrm{L}} \leq f_{ij}^* \leq f_{ij}^{\mathrm{R}}$ and so $\{t_{ij}^{\mathrm{EW}}\}$ is a feasible solution of $\mathcal{P}^*$. The guarantee for $\mathrm{TPR}_{\pi^{\mathrm{EW}}}(X)$ can be established based on a fine-tuned analysis of Lemma 3. See the Appendix for the formal proof.

4.4. Minimax Lower Bound

In this final subsection, we provide a sketch for showing Proposition 1 based on the hypothesis testing argument. We consider the following special model: let $p_{\mathcal{O}} = 1$ and $p_{\mathrm{A}}^* = \frac{1}{2}$, and when $B_{ij} = 1$, $X_{ij} = 0$. We refer to this in notational form as $X \sim \mathrm{H}(M^*)$.

We construct $\mathcal{M}_n = \{M^b \in \mathbb{R}^{n\times n}, b \in \{0,1\}^{n/2}\}$ as follows. Fix a constant C. Consider $b \in \{0,1\}^{n/2}$. For any $i \in [n/2], j \in [n]$, if $b_i = 0$, $M_{2i,j}^b = 1$ and $M_{2i+1,j}^b = 1 - \frac{C}{\sqrt{n}}$; if $b_i = 1$, $M_{2i,j}^b = 1 - \frac{C}{\sqrt{n}}$ and $M_{2i+1,j}^b = 1$. Let M^b be drawn uniformly form $\mathcal{M}_n$ and $X \sim \mathrm{H}(M^b)$. In order to achieve high TPR, one needs to identify the correct b given X. By our construction, the error probability for distinguishing M^{b_1} and M^{b_2} is large when b_1 is 1-bit different from b_2. This provides a lower bound on identifying b and leads to a $O(1/\sqrt{n})$ regret on TPR. The formal proof can be found in the Appendix. One can also verify that Theorem 1 guarantees $O(\log^{1.5} n/\sqrt{n})$

regret for every $M^b \in \mathcal{M}_n$. This shows that our algorithm is optimal for the TPR up to logarithmic factors.

5. Experiments

To evaluate the empirical performance of the EW algorithm, we first consider a synthetic setting where we compare the performances of our EW algorithm with various state-of-the-art approaches. We then measure performance on real-world data from a large retailer.

5.1. Synthetic Data

We generated an ensemble of 1000 matrices M^* of size $n = m = 100$. The varying parameters of the ensemble include (i) r: the rank of the matrix; (ii) $\bar{M}^* = \frac{1}{nm}\sum_{ij} M^*_{ij}$: the average value of all entries; (iii) $p_{\mathcal{O}}$: the probability of an entry being observed; (iv) p^*_{A}: the probability of an entry where an anomaly occurs; and (v) α^*: the anomaly parameter. When an anomaly occurs, $\mathbb{E}\left(\mathrm{Anom}(\alpha^*, M)\right) = \alpha^* M$.

The parameters were sampled uniformly: $r \in [1, 10], \bar{M}^* \in [1, 10], p_{\mathcal{O}} \in [0.5, 1], p^*_{\mathrm{A}} \in [0, 0.3]$ and $\alpha^* \in [0, 1]$. Each instance was generated in two steps: (i) Generate M^*: for a given choice of r and entrywise 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 $\bar{M}^* = \frac{1}{nm}\sum_{ij} M^*_{ij}$. This is a typical way of generating M^* with rank r and non-negative entries (Cemgil, 2008). (ii) Observation: If (i, j) is observed, then with probability $1 - p^*_{\mathrm{A}}$, $X_{ij} \sim \mathrm{Poisson}(M_{ij})$; otherwise, $X_{ij} \sim \mathrm{Poisson}(\mathrm{Exp}(\alpha^*)M_{ij})$. Here $\mathrm{Exp}(\alpha^*)$ models the occurring time of the anomalous event.

We compared our EW algorithm with three existing algorithms: (i) Stable-PCP (Zhou et al., 2010; Chen et al., 2020b), (ii) Robust Matrix Completion (RMC) (Klopp et al., 2017), and (iii) Direct Robust Matrix Factorization (DRMF) (Xiong et al., 2011). These three algorithms all recover the matrices by decomposing $X = \hat{M} + \hat{A} + \hat{E}$ and minimizing $f(\hat{M}) + \lambda_1 g(\hat{A}) + \lambda_2 h(\hat{E})$ where f, g, h are penalty functions with Lagrange multipliers λ_1, λ_2. For all algorithms, we tuned Lagrange multipliers corresponding to rank using knowledge of the true rank. In order to generate ROC curves and compute AUCs, we varied γ in our EW algorithm. For three existing optimization algorithms, we do this by varying the Lagrange multipliers.

The results are summarized in Table 2 and Figure 1. Table 2 reports AUC, $\|\hat{M} - M\|_{\mathrm{F}}$, and $\|\hat{M} - M\|_{\max}$ averaged over 1000 instances for four algorithms ($\hat{M}$ of EW is obtained after recovering from the estimated scaling). The results show that EW achieves an AUC close to π^* (the ideal algorithm that knows M^* and the anomaly model), confirming Theorem 1. For all considered metrics, the results also demonstrate that EW outperforms other algorithms significantly. Figure 1 (Left) shows that the above phenomenon holds uniformly over the ensemble.[4] Figure 1 (Right) shows the explicit ROC curve for a representative setting.

Algorithm	AUC	$\|\hat{M} - M\|_{\mathrm{F}}$	$\|\hat{M} - M\|_{\max}$
π^*	0.823	–	–
EW	0.803	237.1	27.4
Stable PCP	0.708	314.3	43.6
DRMF	0.549	391.2	60.4
RMC	0.519	1099.0	123.1

Table 2: Summary of results on synthetic data. AUC, $\|\hat{M} - M\|_{\mathrm{F}}$, and $\|\hat{M} - M\|_{\max}$ are averaged over 1000 instances. Evaluated algorithms include an ideal algorithm that knows M^* and the anomaly model (π^*), our algorithm (EW), and three existing benchmarks.

5.2. Real Data

We collected data, from a retailer, consisting of weekly sales of $m = 290$ products across $n = 2481$ stores with $p_{\mathcal{O}} \sim 0.14$ and mean value 2.64. Since there is no ground-truth for anomalies, we viewed the collected data as the underlying matrix M^*, and then introduced noise and artificial anomalies. Specifically, we generated X as in the synthetic data (with fixed M^*), introducing anomalies by deliberately perturbing a fraction p^*_{A} of entries and thinning the resulted sales at rate α^*. In particular, for each sample, $p^*_{\mathrm{A}} \in [0, 0.3], \alpha^* \in [0, 1]$ were uniformly drawn. We generated an ensemble of 1000 such perturbed matrices.

Figure 2 reports the results. 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. In particular, the average AUC of π^* is 0.733, the average AUC of π^{EW} is 0.672, whereas the average AUC of Stable-PCP is 0.566. Right of the Figure 2 shows an ROC curve for a representative setting of $p^*_{\mathrm{A}} = 0.04$ and $\alpha^* = 0.2$ [5] where we see the absolute performance. Our results also confirm the EW's ability to recover M^* for real-world data and simulated anomalies. We left the experiments on real anomalies as the future work. More details about experiments can be found in the Appendix.

Scalability: EW is also much faster than compared algorithms, since our main computational cost is a typical matrix completion procedure (our linear program step can be solved via a sort). Concretely, we can expect to solve a 70000×10000 matrix with 10^7 observed entries within

[4] We show results vs. Stable-PCP, but the same holds true for the other two existing algorithms in the experiments.

[5] The parameters are chosen to fit the reported loss caused by the phantom inventory (Gruen et al., 2002).

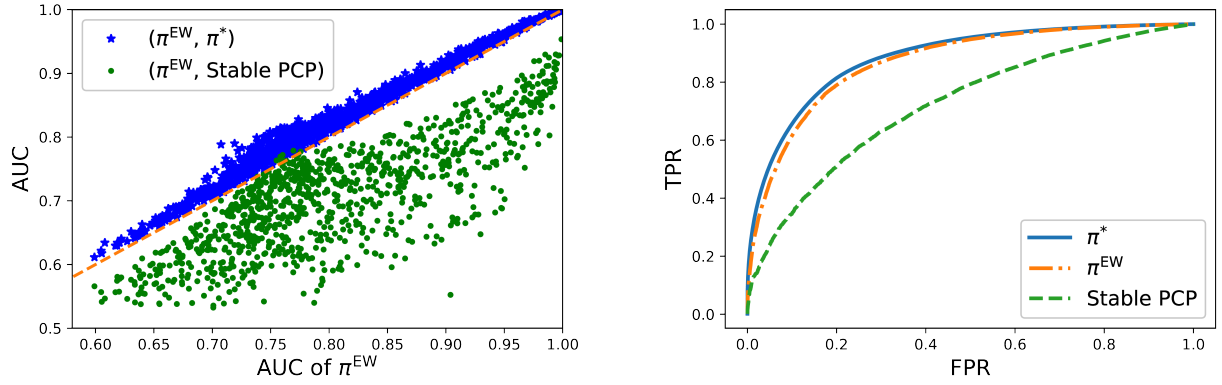

Figure 1: **Synthetic data.** (Left) Scatter plot showing AUC of ideal algorithm vs. that of EW (blue points, above 45-degree line); and AUC of Stable PCP vs EW (green, mostly below 45 degree line). (Right) ROC curve in a representative setting with $n = m = 100, r = 3, \bar{M}^* = 5, p_{\mathcal{O}} = 0.8, p^*_{\mathrm{A}} = 0.04, \alpha^* = 0.2$.

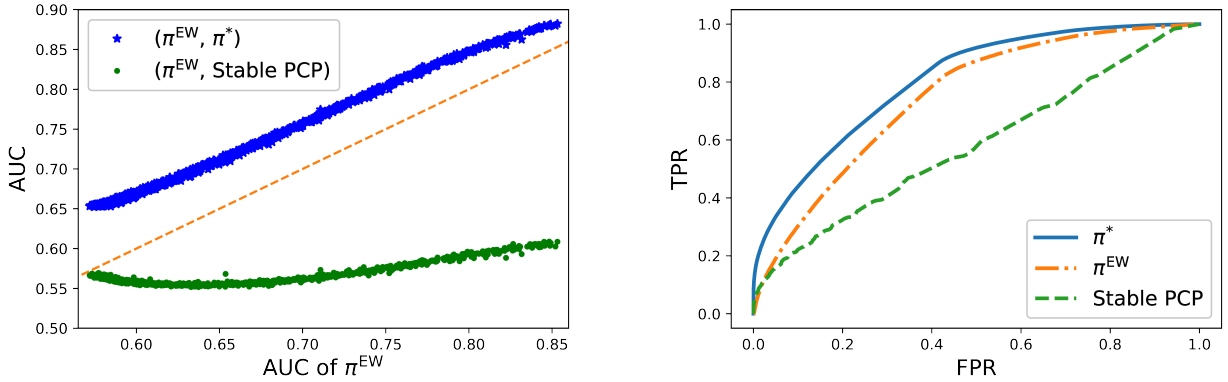

Figure 2: **Real data.** (Left) An ensemble similar to the synthetic data. (Right) ROC curve in a representative setting of $p^*_{\mathrm{A}} = 0.04$ and $\alpha^* = 0.2$.

minutes (Yao & Kwok, 2018).

6. Conclusion

We proposed a simple statistical model for anomaly detection in low-rank matrices that is motivated by the phantom inventory problem in retail. We proved a new entrywise bound for matrix completion with sub-exponential noise, and used this to motivate a simple policy for the anomaly detection problem. We proved matching upper and lower bounds on the anomaly detection rate of our algorithm, and demonstrated in experiments that our approach provides substantial improvements over existing approaches.

While our results are somewhat encouraging, they by no means cover the most general settings of practical interest. There are many possible extensions that merit future investigation, to name a few,

- *Less Restrictive Ω and B.* Our current model for Ω or B is dedicated to a uniformly random model that has been widely used in the literature since (Candès et al., 2011). We think it is an exciting direction to study the noisy anomaly detection problem with less restrictive Ω and B given the recent advancements in the topic of matrix completion with deterministic missing patterns, e.g., (Chatterjee, 2020).
- *Dependency on K and L.* Our current TPR regret likely scales sub-optimal with K amd L. A more refined analysis may lead to the improvement for such dependency.

In addition, as our real dataset validated the core of the theory (e.g., the algorithm ability to recover M^*, which is crucial to separate S), it is promising to test our algorithm for data with ground-truth anomalies and we leave this for future work.

7. Acknowledgement

We thank all anonymous reviewers for their constructive comments. Farias and Peng were partially supported by NSF Grant CMMI 1727239.

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

Agarwal, A., Negahban, S., Wainwright, M. J., et al. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2): 1171–1197, 2012.

Bernstein, S. The theory of probabilities, 1946.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3): 1–37, 2011.

Cao, Y. and Xie, Y. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2015.

Cemgil, A. T. Bayesian inference for nonnegative matrix factorisation models. *Computational intelligence and neuroscience*, 2009, 2008.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Chatterjee, S. A deterministic theory of low rank matrix completion. *IEEE Transactions on Information Theory*, 66(12):8046–8055, 2020.

Chen, J., Liu, D., and Li, X. Nonconvex rectangular matrix completion via gradient descent without $l_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 2020a.

Chen, L. and Mersereau, A. J. Analytics for operational visibility in the retail store: The cases of censored demand and inventory record inaccuracy. In *Retail supply chain management*, pp. 79–112. Springer, 2015.

Chen, Y., Fan, J., Ma, C., and Yan, Y. Inference and uncertainty quantification for noisy matrix completion. *arXiv preprint arXiv:1906.04159*, 2019.

Chen, Y., Fan, J., Ma, C., and Yan, Y. Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *arXiv preprint arXiv:2001.05484*, 2020b.

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

DeHoratius, N. and Raman, A. Inventory record inaccuracy: An empirical analysis. *Management science*, 54(4):627–641, 2008.

DeHoratius, N., Mersereau, A. J., and Schrage, L. Retail inventory management when records are inaccurate. *Manufacturing & Service Operations Management*, 10 (2):257–277, 2008.

Fan, T.-J., Chang, X.-Y., Gu, C.-H., Yi, J.-J., and Deng, S. Benefits of rfid technology for reducing inventory shrinkage. *International Journal of Production Economics*, 147: 659–665, 2014.

Gross, D. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Gruen, T. W., Corsten, D. S., and Bharadwaj, S. *Retail out-of-stocks: A worldwide examination of extent, causes and consumer responses*. Grocery Manufacturers of America Washington, DC, 2002.

Hall, A. R. *Generalized method of moments*. Oxford university press, 2005.

Hansen, L. P. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pp. 1029–1054, 1982.

Imbens, G. W., Johnson, P., and Spady, R. H. Information theoretic approaches to inference in moment condition models. Technical report, National Bureau of Economic Research, 1995.

Klopp, O., Lounici, K., and Tsybakov, A. B. Robust matrix completion. *Probability Theory and Related Fields*, 169 (1-2):523–564, 2017.

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

Lafond, J. Low rank matrix completion with exponential family noise. In *Conference on Learning Theory*, pp. 1224–1243, 2015.

Ma, C., Wang, K., Chi, Y., and Chen, Y. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, pp. 1–182, 2019.

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

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug): 2287–2322, 2010.

McRae, A. D. and Davenport, M. A. Low-rank matrix completion and denoising under poisson noise. *arXiv preprint arXiv:1907.05325*, 2019.

Nachtmann, H., Waller, M. A., and Rieske, D. W. The impact of point-of-sale data inaccuracy and inventory record data errors. *Journal of Business Logistics*, 31(1): 149–158, 2010.

Newey, K. and McFadden, D. Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pp. 2112–2245, 1994.

Raman, A., DeHoratius, N., and Ton, Z. Execution: The missing link in retail operations. *California Management Review*, 43(3):136–152, 2001.

Sambasivan, A. V. and Haupt, J. D. Minimax lower bounds for noisy matrix completion under sparse factor models. *IEEE Transactions on Information Theory*, 64(5):3274–3285, 2018.

Shi, J., Katehakis, M. N., Melamed, B., and Xia, Y. Production-inventory systems with lost sales and compound poisson demands. *Operations Research*, 62(5): 1048–1063, 2014.

Tropp, J. A. et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Wang, F., Fang, X., Chen, X., and Li, X. Impact of inventory inaccuracies on products with inventory-dependent demand. *International Journal of Production Economics*, 177:118–130, 2016.

Wong, R. K. and Lee, T. C. Matrix completion with noisy entries and outliers. *The Journal of Machine Learning Research*, 18(1):5404–5428, 2017.

Xiong, L., Chen, X., and Schneider, J. Direct robust matrix factorizatoin for anomaly detection. In *2011 IEEE 11th International Conference on Data Mining*, pp. 844–853. IEEE, 2011.

Yao, Q. and Kwok, J. T. Accelerated and inexact soft-impute for large-scale matrix and tensor completion. *IEEE Transactions on Knowledge and Data Engineering*, 31(9):1665–1679, 2018.

Zhou, Z., Li, X., Wright, J., Candes, E., and Ma, Y. Stable principal component pursuit. In *2010 IEEE international symposium on information theory*, pp. 1518–1522. IEEE, 2010.