Noisy Bilinear Low-Rank Matrix Sketching

Wenbin Wang, Xindi Ping, Cheng Cheng, and Ziping Zhao
School of Information Science and Technology, ShanghaiTech University, Shanghai, China
{wangwb2023, pingxd2023, chengcheng2, zipingzhao}@shanghaitech.edu.cn

2025 IEEE Information Theory Workshop (ITW) | 979-8-3315-3142-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ITW62417.2025.11240268

***Abstract*—This paper studies the problem of recovering a low-rank matrix from noisy bilinear measurements, which arises in a range of real-world applications. We propose a novel estimator that minimizes a least-squares loss regularized by a nonconvex penalty to promote low-rank structure. To solve the resulting nonconvex problem, we develop an efficient proximal gradient descent algorithm. We show that, under mild conditions, the proposed estimator consistently recovers the underlying matrix and achieves the statistically optimal convergence rate. Numerical experiments on both synthetic and real-world datasets validate the theoretical guarantees and demonstrate the practical effectiveness of the proposed method.**

***Index Terms*—Compressive sensing, bilinear measurements model, matrix sketching, covariance sketching, graph sketching.**

I. Introduction

Compressive sensing has significantly advanced the field of signal processing by enabling the accurate reconstruction of high-dimensional signals from a reduced number of measurements, by exploiting the inherent sparsity of the signal representation in an appropriate basis [1]–[4]. By relaxing the conventional Nyquist–Shannon sampling rate requirement, compressive sensing allows for sub-Nyquist signal acquisition, which is especially beneficial in applications constrained by limited bandwidth, sampling rate, or energy budget. This framework has found extensive applications in diverse domains, including dynamic background subtraction in video surveillance [5], accelerated magnetic resonance imaging [6], and wideband spectrum sensing in cognitive radio systems [7], as well as in numerous other contexts [8]–[13]. The theory of compressive sensing establishes that a high-dimensional sparse signal vector $\boldsymbol{x} \in \mathbb{R}^d$ can be stably and accurately recovered from an underdetermined system of linear measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times d}$ is a suitably constructed sensing matrix with $m \ll d$, provided that it satisfies certain structural conditions such as mutual incoherence or the restricted isometry property [14]–[16]. It has been theoretically established that $m = \mathcal{O}(s\log(d/s))$ measurements are sufficient to recover an s-sparse signal when the sensing matrix is drawn from a sub-Gaussian distribution [17], and that $m = \mathcal{O}(s\log^2(d/s))$ measurements suffice when the sensing matrix follows a sub-exponential distribution [18].

Recent advances have extended the classical compressive sensing paradigm from sparse vector recovery to the recovery of structured matrices from compressed observations [19]–[23]. In particular, one class of problems considers recovering low-rank matrices from linear measurements [22]. However, in many practical applications, the measurements are bilinear in nature [21]. A canonical bilinear measurement model is given by

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}^\star\boldsymbol{B}^\top + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times d}$ are sketching (or sensing) matrices, $\boldsymbol{X}^\star \in \mathbb{R}^{d\times d}$ is a low-rank matrix, and $\boldsymbol{E}$ denotes a possible additive noise.

The model (1) arises naturally in various applications, including graph sketching [24], [25] and covariance sketching [20], [26]. In the context of graph sketching, consider a large-scale graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the size of $\mathcal{V}$ poses significant challenges for storage, communication, and computation. A common strategy is to construct a compressed graph $\mathcal{G}'$ by partitioning the vertex set $\mathcal{V}$ into m disjoint subsets $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m$. In $\mathcal{G}'$, each partition $\mathcal{V}_i$ is represented by a single node, and the weight of the edge between nodes $\mathcal{V}_i$ and $\mathcal{V}_j$ corresponds to the total number of edges in $\mathcal{G}$ connecting vertices in $\mathcal{V}_i$ to those in $\mathcal{V}_j$. Let $\boldsymbol{A}_i$ denote the indicator row vector for partition $\mathcal{V}_i$, i.e., $(\boldsymbol{A}_i)_j = 1$ if node $j \in \mathcal{V}_i$, and 0 otherwise. Stacking these vectors yields a matrix $\boldsymbol{A}$, and the adjacency matrix of the compressed graph is then given by $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}^\star\boldsymbol{A}^\top$, where $\boldsymbol{X}^\star$ is the adjacency matrix of the original graph $\mathcal{G}$. In covariance sketching, consider two zero-mean random vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ with an unknown cross-covariance matrix $\boldsymbol{X}^\star = \mathrm{E}(\boldsymbol{x}\boldsymbol{x}'^\top)$. One applies sketching matrices $\boldsymbol{A}, \boldsymbol{B}$ and observes the compressed vectors $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{z}' = \boldsymbol{B}\boldsymbol{x}'$. The cross-covariance matrix of the sketches then satisfies $\mathrm{E}(\boldsymbol{z}\boldsymbol{z}'^\top) = \boldsymbol{A}\boldsymbol{X}^\star\boldsymbol{B}^\top$. In practice, however, one typically relies on the empirical cross-covariance of finite samples, which introduces a noise component $\boldsymbol{E}$, yielding the model $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}^\star\boldsymbol{B}^\top + \boldsymbol{E}$. In [21], the problem of noiseless recovery is studied under the assumption that the target matrix $\boldsymbol{X}^\star$ exhibits distributed sparsity. The proposed recovery approach involves solving the following convex optimization problem:

$$\begin{aligned} \underset{\boldsymbol{X}}{\text{minimize}} \quad & \|\boldsymbol{X}\|_1 \\ \text{subject to} \quad & \boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{B}^\top. \end{aligned} \tag{2}$$

It is shown that exact recovery is achievable with high probability when the sketching dimension satisfies $m = \mathcal{O}(\sqrt{sd}\log d)$ in the absence of noise. However, extending

The experiments of this work were supported by the core facility Platform of Computer Science and Communication, SIST, ShanghaiTech University.

979-8-3315-3142-3/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: ShanghaiTech University. Downloaded on January 20,2026 at 15:39:03 UTC from IEEE Xplore. Restrictions apply.

this analysis to the noisy case and developing estimators with improved statistical guarantees remain important open problems.

In addition to sparsity, alternative structural priors on $\boldsymbol{X}^\star$, such as low-rankness, are frequently encountered and are crucial in many practical scenarios, including phase retrieval [27], power spectrum estimation [28], collaborative filtering [29], and metric learning [30]. Advances in the theory and algorithms for sparse signal recovery provide valuable insights into low-rank matrix recovery, since low-rankness can be interpreted as sparsity in the singular values of a matrix. Classical methods such as basis pursuit [31], orthogonal matching pursuit [32], and the iterative shrinkage-thresholding algorithm [33] have laid a strong foundation in the sparse recovery literature. Among these, the least absolute shrinkage and selection operator (LASSO) [34] stands out due to its computational tractability and well-established statistical guarantees. However, LASSO may introduce estimation bias and fail to achieve consistent support recovery, particularly in the presence of highly correlated variables [35], [36]. To address these limitations, nonconvex regularization methods have been developed as tighter approximations to the ideal ℓ_0 penalty. Prominent examples include the smoothly clipped absolute deviation (SCAD) [37] and the minimax concave penalty (MCP) [38], which provide improved support recovery and reduced estimation bias. Since the nuclear norm is essentially an ℓ_1 norm applied to the singular values, these nonconvex penalties naturally extend to low-rank matrix recovery.

In this paper, we propose a novel estimation method for low-rank matrix recovery from noisy bilinear measurements, employing a nonconvex penalty on the singular values to promote low-rank structure. To solve the resulting nonconvex optimization problem, we develop an efficient algorithm based on a proximal gradient homotopy method. Furthermore, we rigorously establish that the proposed estimator possesses an oracle property under a minimal signal strength condition, guaranteeing exact recovery of the true rank of the underlying matrix.

II. Proposed Method

In this section, we develop a nonconvex regularized estimator for low-rank matrix recovery under bilinear measurements. Then, to compute the proposed estimator, we develop an efficient optimization algorithm based on proximal gradient.

A. Proposed Estimator Based on Nonconvex Penalty

To estimate a low-rank matrix $\boldsymbol{X}$ from observed data $\boldsymbol{Y}$ with bilinear noisy measurements, we consider a nonconvex regularized least-squares optimization framework. Specifically, we introduce the following estimator:

$$\underset{\boldsymbol{X}}{\text{minimize}} \ \frac{1}{2m^2}\left\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}^\top\right\|_{\mathrm{F}}^2 + P_\lambda(\boldsymbol{X}), \tag{3}$$

where $P_\lambda(\boldsymbol{X}) = \sum_{i=1}^d p_\lambda(\sigma_i(\boldsymbol{X}))$ is a decomposable nonconvex penalty imposed on the singular values of $\boldsymbol{X}$, governed by a tuning parameter $\lambda > 0$. Such a regularization structure leverages sparsity in singular values, promoting low-rank solutions. The penalty term is characterized by a decomposition:

$$P_\lambda(\boldsymbol{X}) = \lambda\|\boldsymbol{X}\|_* + Q_\lambda(\boldsymbol{X}),$$

where $\|\boldsymbol{X}\|_*$ denotes the nuclear norm and $Q_\lambda(\boldsymbol{X}) = \sum_{i=1}^d q_\lambda(\sigma_i(\boldsymbol{X}))$ encapsulates a concave adjustment applied to individual singular values $\sigma_i(\boldsymbol{X})$. Correspondingly, the scalar penalty function $p_\lambda(t)$ admits a decomposition into a standard ℓ_1 penalty and a concave perturbation $p_\lambda(t) = \lambda|t| + q_\lambda(t)$, where $q_\lambda(t)$ introduces nonconvexity.

Two prominent examples within this penalty framework include SCAD [37] and MCP [38].

Assumption 1. *The penalty functions $p_\lambda(t)$ and the associated concave components $q_\lambda(t)$ satisfy the following properties:*

1) *There exists $\nu > 0$ such that the derivative satisfies $p'_\lambda(t) = 0$ for all $t \geq \nu$;*
2) *Both $p_\lambda(t)$ and $q_\lambda(t)$ are symmetric about zero, i.e., $p_\lambda(t) = p_\lambda(-t)$, $q_\lambda(t) = q_\lambda(-t)$;*
3) *The derivative $q'_\lambda(t)$ is monotonic and Lipschitz continuous in the interval $[0, \infty)$. Explicitly, for $t_2 \geq t_1 \geq 0$, there exist constants $\zeta^- \geq \zeta^+ > 0$ such that $-\zeta^- \leq \frac{q'_\lambda(t_2)-q'_\lambda(t_1)}{(t_2-t_1)} \leq -\zeta^+$.*
4) *Both $q_\lambda(t)$ and its derivative vanish at zero, i.e., $q_\lambda(0) = q'_\lambda(0) = 0$;*
5) *There exists a constant $\lambda > 0$ bounding the magnitude of the derivative, i.e., $|q'_\lambda(t)| \leq \lambda$.*

These conditions highlight essential structural characteristics. Particularly, condition (3) emphasizes the concavity level, directly influencing the degree of nonconvexity of the penalty function.

B. Optimization Algorithm

Define $F(\boldsymbol{X}) = \tilde{f}(\boldsymbol{X}) + \lambda\|\boldsymbol{X}\|_*$, where $\tilde{f}(\boldsymbol{X}) = f(\boldsymbol{X}) + Q_\lambda(\boldsymbol{X})$ and $f(\boldsymbol{X}) = \frac{1}{2m^2}\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\boldsymbol{B}^\top\|_{\mathrm{F}}^2$. The proposed methodology operates by iteratively constructing a quadratic surrogate function to approximate the function $\tilde{f}(\boldsymbol{X})$ around the current estimate, thereby yielding a locally convex subproblem. The regularization parameter λ is progressively decreased throughout iterations, leveraging the homotopy continuation approach to facilitate convergence towards a global solution.

Quadratic Approximation: A second-order approximation of $\tilde{f}(\boldsymbol{X})$ around the point $\boldsymbol{M}$ is given by:

$$\begin{aligned}\widetilde{F}(\boldsymbol{X};\boldsymbol{M}) =& \tilde{f}(\boldsymbol{M}) + \langle\nabla\tilde{f}(\boldsymbol{M}), \boldsymbol{X}-\boldsymbol{M}\rangle + \frac{L}{2}\|\boldsymbol{X}-\boldsymbol{M}\|_{\mathrm{F}}^2 \\ &+ \lambda\|\boldsymbol{X}\|_*,\end{aligned} \tag{4}$$

where L is a Lipschitz constant of $\tilde{f}(\boldsymbol{X})$ and $\langle\cdot,\cdot\rangle$ is the inner product. In practice, we initialize the Lipschitz constant with a conservative lower bound $L_{\min}$ and iteratively adjust this estimate by a multiplicative factor (typically doubling $L_{\min}$) until suitable convergence criteria are met. Consequently, the

Authorized licensed use limited to: ShanghaiTech University. Downloaded on January 20,2026 at 15:39:03 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Proximal Gradient Algorithm

```
Input: $\lambda_0 > 0$, $\epsilon > 0$, $L_{\min} > 0$, $\eta \in (0,1)$, $\delta \in (0,1)$
Initialize: $\boldsymbol{X}^0 = \boldsymbol{0}$, $L_0 = L_{\min}$
for $t = 0, 1, \ldots, T-1$ do
    $\lambda_{t+1} = \eta\lambda_t$;
    $\epsilon_{t+1} = \lambda_t/4$;
    $k = 0$;
    $\boldsymbol{X}^k = \boldsymbol{X}^t$;
    while $\omega_{\lambda_{t+1}}(\boldsymbol{X}^k) > \epsilon_{t+1}$ do
        $k = k + 1$;
        $\boldsymbol{X}^k = \arg\min_{\boldsymbol{X}} \widetilde{F}_{L,\lambda}(\boldsymbol{X}; \boldsymbol{X}^{k-1})$;
        if $F(\boldsymbol{X}^k) > \widetilde{F}(\boldsymbol{X}^k; \boldsymbol{X}^{k-1})$ then
            $L_{k-1} = 2L_{k-1}$
        end
        $L_k = \max\{L_{\min}, L_{k-1}/2\}$;
    end
    $\boldsymbol{X}^{t+1} = \boldsymbol{X}^k$;
    $L_{t+1} = L_k$;
end
Output: $\{\boldsymbol{X}^t\}_{t=1}^T$
```

optimization problem reduces to finding the minimizer of the surrogate function $\widetilde{F}(\boldsymbol{X}; \boldsymbol{M})$:

$$\boldsymbol{X} \in \arg\min_{\boldsymbol{X}} \widetilde{F}(\boldsymbol{X}; \boldsymbol{M}) \tag{5}$$

The minimization problem in (5) can be efficiently solved using the singular value thresholding method, as described in [39], [40].

Optimality Conditions and Duality: Let $\widehat{\boldsymbol{X}}$ be the global minimizer of the optimization problem (3). The optimality condition for $\widehat{\boldsymbol{X}}$ can be characterized by

$$\langle \widehat{\boldsymbol{X}} - \boldsymbol{X}, \nabla \tilde{f}(\widehat{\boldsymbol{X}}) + \lambda \boldsymbol{\Upsilon}' \rangle \leq 0, \tag{6}$$

where $\boldsymbol{\Upsilon} \in \partial \|\widehat{\boldsymbol{X}}\|_*$ is a subgradient of the nuclear norm at $\widehat{\boldsymbol{X}}$. Since an analytical solution does not exist, the exact solution can never be achieved. To quantify optimality, we consider an approximate solution. First, we introduce the duality gap, $\omega_\lambda(\boldsymbol{X})$, which measures the suboptimality of a matrix $\boldsymbol{X}$:

$$\begin{aligned}\omega_\lambda(\boldsymbol{X}) &= \min_{\boldsymbol{\Upsilon}' \in \partial \|\widehat{\boldsymbol{X}}\|} \max_{\boldsymbol{X}'} \left\{ \frac{\langle \widehat{\boldsymbol{X}} - \boldsymbol{X}, \nabla \tilde{f}(\widehat{\boldsymbol{X}}) + \lambda \boldsymbol{\Upsilon}' \rangle}{\|\boldsymbol{X} - \boldsymbol{X}'\|_*} \right\} \\ &= \min_{\boldsymbol{\Upsilon}' \in \partial \|\widehat{\boldsymbol{X}}\|} \{ \|\nabla \tilde{f}(\widehat{\boldsymbol{X}}) + \lambda \boldsymbol{\Upsilon}'\|_{\mathrm{F}} \},\end{aligned} \tag{7}$$

In view of the duality gap, if $\boldsymbol{X}$ is the exact minimizer, then $\omega_\lambda(\boldsymbol{X}) \leq 0$. Otherwise, if $\boldsymbol{X}$ is close to the optimum, $\omega_\lambda(\boldsymbol{X})$ will likely be a small positive value.

Regularization Parameter Update: Let the initial regularization parameter be denoted as $\lambda = \lambda_0$, where λ_0 is chosen to be sufficiently large. The parameter λ is progressively decreased at each iteration following the formula: $\lambda_t = \eta^t \lambda_0$, where η is a constant. To guarantee that each iteration step attains the prescribed accuracy, we further force $\epsilon_t < \lambda_t/4$.

III. Main Results

In this section, we provide formal theoretical results for our proposed estimator. We begin with some necessary preliminaries.

Consider the ground truth matrix $\boldsymbol{X}^\star$ with singular value decomposition (SVD) given by $\boldsymbol{X}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$, where $\boldsymbol{U}^\star, \boldsymbol{V}^\star \in \mathbb{R}^{d \times r}$, and $\boldsymbol{\Sigma}^\star = \operatorname{diag}(\sigma_1^\star, \ldots, \sigma_r^\star)$. We introduce the subspace $\mathcal{F}$ and $\mathcal{F}^\perp$, which are defined in terms of the row and column spaces of the matrices:

$$\begin{aligned}\mathcal{F}(\boldsymbol{U}^\star, \boldsymbol{V}^\star) &:= \{ \boldsymbol{\Delta} \mid \mathrm{row}(\boldsymbol{\Delta}) \subseteq \boldsymbol{V}^\star, \mathrm{col}(\boldsymbol{\Delta}) \subseteq \boldsymbol{U}^\star \}, \\ \mathcal{F}^\perp(\boldsymbol{U}^\star, \boldsymbol{V}^\star) &:= \{ \boldsymbol{\Delta} \mid \mathrm{row}(\boldsymbol{\Delta}) \perp \boldsymbol{V}^\star, \mathrm{col}(\boldsymbol{\Delta}) \perp \boldsymbol{U}^\star \}.\end{aligned}$$

Next, we first introduce a restricted set and state two assumptions on the function $f(\boldsymbol{X})$ therein. These conditions have been extensively investigated in prior studies [41], [42], which ensures that $f(\boldsymbol{X})$ is well-conditioned in the local region.

Definition 2. *Define a local region $\mathcal{R}$ as*

$$\mathcal{R} = \{ \boldsymbol{\Delta} \mid \|\Pi_{\mathcal{F}^\perp}(\boldsymbol{\Delta})\|_* \leq 5 \|\Pi_{\mathcal{F}}(\boldsymbol{\Delta})\|_* \}, \tag{8}$$

where $\Pi_{\mathcal{F}(\cdot)}$ is the projection operator that projects matrices into the subspace $\mathcal{F}$.

Assumption 3. *The empirical loss function $f(\cdot)$ is ρ^--strongly convex and ρ^+-smooth over $\mathcal{R}$ with $\infty > \rho^+ \geq \rho^- > 0$. Specifically, for all $\boldsymbol{X} - \boldsymbol{X}' \in \mathcal{C}$, we have:*

$$\left\langle \boldsymbol{X} - \boldsymbol{X}', \nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}') \right\rangle \geq \rho^- \|\boldsymbol{X} - \boldsymbol{X}'\|_{\mathrm{F}}^2,$$

$$\left\langle \boldsymbol{X} - \boldsymbol{X}', \nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}') \right\rangle \geq \frac{\|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}')\|_{\mathrm{F}}^2}{\rho^+}.$$

These conditions are standard structural assumptions ensuring desirable curvature properties of $f(\boldsymbol{X})$. It has been demonstrated that, with high probability, $f(\boldsymbol{X})$ can be shown to meet these conditions [43].

We now proceed to establish a deterministic bound.

Theorem 4. *Define $\mathcal{S}_1 = \{ i \mid \sigma_i^\star \geq \nu \}$, $\mathcal{S}_1 = \{ i \mid \nu > \sigma_i^\star > 0 \}$ with their corresponding cardinalities given by $s_1 = |\mathcal{S}_1|$ and $s_2 = |\mathcal{S}_2|$. Suppose Assumptions 1 and 3 hold, if $\rho^- > \zeta^-$, $\lambda \gtrsim \|\boldsymbol{A}^\top \boldsymbol{E} \boldsymbol{B}\|_{\mathrm{F}} / m^2$, we have:*

$$\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_{\mathrm{F}} \lesssim \tau \sqrt{s_1} + \sqrt{s_2} \tag{9}$$

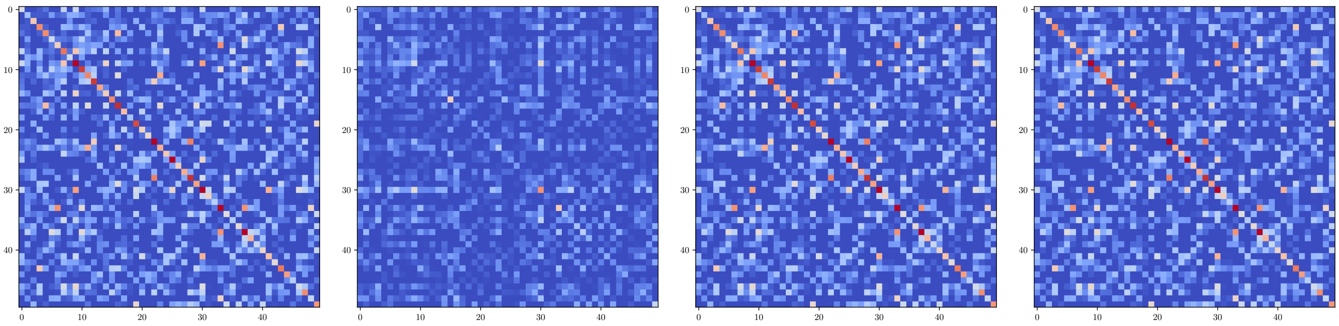
where $\tau = \|\Pi_{\mathcal{F}_{S_1}}(\nabla f(\boldsymbol{X}^\star))\|_{\mathrm{F}}$ and $\mathcal{F}_{S_1}$ is a subspace of $\mathcal{F}$ associated with S_1.

Note that the derived error bound comprises two distinct components: one corresponding to singular values of large magnitude, and the other corresponding to those of smaller magnitude. In what follows, we introduce the oracle estimator, which serves as an essential benchmark for assessing the performance of the proposed estimator.

Remark 5. *The oracle rate refers to the statistical convergence rate of the oracle estimator, which knows the true rank subspace $\mathcal{F}(\boldsymbol{U}^\star, \boldsymbol{V}^\star)$. The oracle estimator $\widehat{\boldsymbol{X}}^O$ is defined as*

$$\widehat{\boldsymbol{X}}^O = \arg\min_{\boldsymbol{X} \in \mathcal{F}(\boldsymbol{U}^\star, \boldsymbol{V}^\star)} f(\boldsymbol{X})$$

Authorized licensed use limited to: ShanghaiTech University. Downloaded on January 20,2026 at 15:39:03 UTC from IEEE Xplore. Restrictions apply.



(a) Ground Truth (b) Nuclear (c) MCP (d) SCAD

Fig. 1: Visualization of the ground truth and reconstructed estimates obtained under different regularization penalties.

According to the definition, it is easy to obtain that $\widehat{\boldsymbol{X}}^{O}$ satisfies $\|\widehat{\boldsymbol{X}}^{O}-\boldsymbol{X}^{\star}\|_{\mathrm{F}} \lesssim \|\Pi_{\mathcal{F}}\left(\nabla f\left(\boldsymbol{X}^{\star}\right)\right)\|_{\mathrm{F}}$.

It is worth noting that the first term on the RHS of (9) closely approaches the oracle convergence rate. Furthermore, this term becomes dominant provided that the number of singular values with small magnitudes remains sufficiently limited. Therefore, we introduce the following minimum signal strength condition.

Assumption 6. *The singular value of the ground truth $\boldsymbol{X}^{\star}$ satisfies*

$$\min_{i\in\mathcal{S}_1\cup\mathcal{S}_2}|\sigma_i^{\star}| \geq \nu + 2\sqrt{s_1+s_2}\|\boldsymbol{A}^{\top}\boldsymbol{E}\boldsymbol{B}\|_{\mathrm{F}}/(m^2\rho). \tag{10}$$

Assumption 6 is commonly adopted in the analysis of nonconvex penalized regression problems [22], [44], [45]. This condition is relatively mild since the second term tends to diminish significantly as the sample size m increases. Consequently, Assumption 6 ensures that the set $\mathcal{S}_2$ becomes empty. Next, we give the oracle property of our estimator.

Theorem 7 (Oracle Property)**.** *Suppose Assumptions 1, 3 and 6 hold. If $\rho > \zeta^{-}$, and $\lambda \geq \frac{(\rho^{-}+\sqrt{s_1+s_2}\rho^{+})\|\boldsymbol{A}^{\top}\boldsymbol{E}\boldsymbol{B}\|_{\mathrm{F}}}{2m^2\rho^{-}}$, we have $\operatorname{rank}(\widehat{\boldsymbol{X}}) = \operatorname{rank}(\widehat{\boldsymbol{X}}^{O}) = \operatorname{rank}\left(\boldsymbol{X}^{\star}\right)$ and*

$$\|\widehat{\boldsymbol{X}}-\boldsymbol{X}^{\star}\|_{\mathrm{F}} \lesssim \sqrt{s_1}\tau, \tag{11}$$

where $\tau = \|\Pi_{\mathcal{F}}\left(\nabla f\left(\boldsymbol{X}^{\star}\right)\right)\|_{\mathrm{F}}$.

Next, we consider a more practical scenario where the noise entries are independently and identically distributed sub-Gaussian random variables with variance κ and the vectorized sketching matrices $\operatorname{vec}(\boldsymbol{A}), \operatorname{vec}(\boldsymbol{B}) \in \mathbb{R}^{m^2}$ follow sub-Gaussian distributions. That is $\operatorname{vec}(\boldsymbol{A}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Theta}_1)$, $\operatorname{vec}(\boldsymbol{B}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Theta}_2)$. We term these sensing matrices as $\boldsymbol{\Theta}_1$-ensemble and $\boldsymbol{\Theta}_2$-ensemble, respectively. Define $\varpi_1(\boldsymbol{\Theta}_1) = \sqrt{\sup_{\|\boldsymbol{u}\|_2=1,\|\boldsymbol{v}\|_2=1}\operatorname{Var}(\boldsymbol{u}^{\top}\boldsymbol{A}\boldsymbol{v})}$ and $\varpi_2(\boldsymbol{\Theta}_2) = \sqrt{\sup_{\|\boldsymbol{u}\|_2=1,\|\boldsymbol{v}\|_2=1}\operatorname{Var}(\boldsymbol{u}^{\top}\boldsymbol{B}\boldsymbol{v})}$. Then, we provide explicit statistical guarantees under the sub-Gaussian design.

Corollary 8. *Suppose Assumptions 1 and 3 hold. Consider the random design matrices $\boldsymbol{A}$ sampled from the $\boldsymbol{\Theta}_1$-ensemble and $\boldsymbol{B}$ sampled from the $\boldsymbol{\Theta}_2$-ensemble. If $\rho \asymp \sqrt{\lambda_{\min}(\boldsymbol{\Theta}_1)\lambda_{\min}(\boldsymbol{\Theta}_2)} > \zeta^{-}$, where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue, and $\lambda \gtrsim \kappa\sqrt{\varpi_1(\boldsymbol{\Theta_1})\varpi_2(\boldsymbol{\Theta_2})d}/m$, then with probability at least $1-\exp(-d)$,*

$$\left\|\widehat{\boldsymbol{X}}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \lesssim \mathcal{O}\left(\sqrt{\frac{\varpi_1(\boldsymbol{\Theta_1})\varpi_2(\boldsymbol{\Theta_2})}{\lambda_{\min}(\boldsymbol{\Theta_1})\lambda_{\min}(\boldsymbol{\Theta_2})}\frac{\kappa}{m}\left[s_1+\sqrt{s_2 d}\right]}\right).$$

This result is a direct consequence of Theorem 4. A closely related problem has been previously addressed via convex relaxation techniques. Specifically, building upon the analysis presented in [46], we have

$$\|\widehat{\boldsymbol{X}}-\boldsymbol{X}^{\star}\|_{\mathrm{F}} \lesssim \mathcal{O}\left(\sqrt{\frac{\varpi_1(\boldsymbol{\Theta_1})\varpi_2(\boldsymbol{\Theta_2})}{\lambda_{\min}(\boldsymbol{\Theta_1})\lambda_{\min}(\boldsymbol{\Theta_2})}\frac{\kappa\sqrt{s_2 d}}{m}}\right).$$

When Assumption 6 is satisfied, one can yield the faster convergence rate $\mathcal{O}\left(\sqrt{\frac{\varpi_1(\boldsymbol{\Theta_1})\varpi_2(\boldsymbol{\Theta_2})}{\lambda_{\min}(\boldsymbol{\Theta_1})\lambda_{\min}(\boldsymbol{\Theta_2})}\frac{\kappa s_1}{m}}\right)$.

IV. Experiments

In this section, we evaluate the empirical performance of our proposed estimator and the corresponding algorithm under both synthetic and real-world settings. For every experimental configuration, we repeat 100 independent trials and report the averaged relative reconstruction error $\frac{\|\widehat{\boldsymbol{X}}-\boldsymbol{X}\|_{\mathrm{F}}}{\|\boldsymbol{X}\|_{\mathrm{F}}}$ over all runs. We employ the SCAD and MCP penalties, where SCAD is given by

$$p_{\lambda}\left(t\right) = \begin{cases} \lambda\left|t\right|, & \text{if } \left|t\right| \leq \lambda, \\ -\frac{t^2-2b\lambda|t|+\lambda^2}{2(b-1)}, & \text{if } \lambda < |t| \leq b\lambda, \\ \frac{(b+1)\lambda^2}{2}, & \text{if } |t| > b\lambda, \end{cases}$$

for some $b > 2$ and MCP is defined as

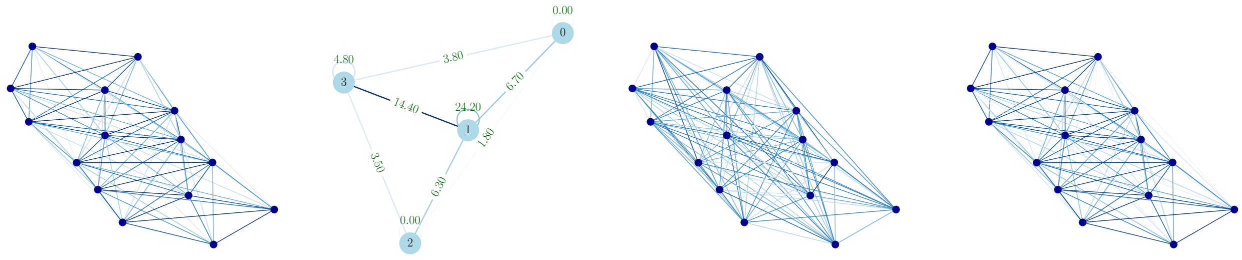
$$p_{\lambda}\left(t\right) = \operatorname{sign}\left(t\right)\lambda \cdot \int_0^{|t|}\left(1-\frac{z}{\lambda b}\right)_{+} dz$$

for some $b > 0$. Throughout, the tuning parameter λ for each method is chosen by five-fold cross-validation, and any additional nonconvex tuning parameter b is selected from a candidate set so as to optimize overall performance.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on January 20,2026 at 15:39:03 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Performance Comparison of Multiple Methods Across Datasets.

Dataset	Nuclear	Weighted Nuclear	SCAD	MCP
Fashion-MNIST [47]	0.7683 ± 0.1076	0.7286 ± 0.0630	0.0124 ± 0.0033	$\mathbf{0.0108 \pm 0.0012}$
Places365 [48]	0.4472 ± 0.0827	0.4647 ± 0.0484	0.0079 ± 0.0013	$\mathbf{0.0066 \pm 0.0021}$
ImageNet-O [49]	0.4574 ± 0.1502	0.5069 ± 0.1149	$\mathbf{0.0137 \pm 0.0079}$	0.0138 ± 0.0056



(a) (b) (c) (d)

Fig. 2: An illustrative example of graph sketching is shown as follows: (a) The original graph $\mathcal{G}$ with 15 nodes; (b) The sketch of the graph $\mathcal{G}$, where the nodes represent the partitions and the edges represent the total number of edges of G that cross these partitions; (c) The graph recovered using least square error minimization; (d) The graph recovered using the SCAD penalty.

A. Synthetic Data

We first evaluate the recovery performances of the nonconvex estimator in (3), instantiated with SCAD and MCP penalties, against the convex nuclear-norm baseline on synthetic data. A ground truth matrix $\boldsymbol{X}^\star \in \mathbb{R}^{50\times 50}$ is generated as $\boldsymbol{X}^\star = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L} \in \mathbb{R}^{50\times 10}$ has independent and identically distributed (i.i.d.) entries drawn from a Gaussian distribution $\mathcal{N}(0,1)$. The sketching matrix $\boldsymbol{A}$ and $\boldsymbol{B}$ are drawn independently with i.i.d. $\mathcal{N}(0,1)$ entries, and the measurement is obtained via model 1 in which each entry of $\boldsymbol{E}$ follows $\mathcal{N}(0,0.01)$. Fig. 1a illustrates the ground-truth low-rank structure. Figs. 1c–1b compare reconstructions from MCP, SCAD, and nuclear-norm minimization, respectively. Both MCP (Fig. 1c) and SCAD (Fig. 1d) yield near-perfect reconstructions. In contrast, the nuclear-norm solution (Fig. 1b) exhibits mild smoothing, which slightly attenuates fine low-rank features. These heatmaps vividly demonstrate the enhanced ability of nonconvex penalties to recover exact low-rank structures, which backs up our theoretical analysis, and the improvement is significant compared with the traditional nuclear-norm penalty. We also evaluate our method on graph sketching. Fig. 2a represents the original graph. Fig. 2b shows the sketched version of the original graph, obtained by applying the sketching matrix $\boldsymbol{A}$, which is randomly generated following a binomial distribution:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Fig. 2c and 2d illustrate the graph recovery results using least squares error minimization and the SCAD penalty function, respectively.

B. Real-world Images

We further validate on three standard image datasets: Fashion-MNIST [47], Places365 [48], and ImageNet-O [49]. A color image is resized and converted to a $d\times d$ grayscale matrix $\boldsymbol{X}$. To enforce low-rank structure, we compute the thin SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and retain only the top r singular components, yielding $\boldsymbol{X}_r = \boldsymbol{U}_{(:,1:r)}\boldsymbol{\Sigma}_{1:r,1:r}\boldsymbol{V}_{(:,1:r)}^\top$. Specifically, we set $(d,r) = (28,10)$ for Fashion-MNIST [47], $(d,r) = (256,100)$ for Places365 [48], and $(d,r) = (512,200)$ for ImageNet-O [49]. We generate two sketching matrices $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{m\times d}$ with i.i.d. entries drawn from $\mathcal{N}(0,1)$, choosing $m = 5, 50, 80$ for Fashion-MNIST [47], Places365 [48], and ImageNet-O [49]. We also generate a noise matrix $\boldsymbol{E}$ whose entries are i.i.d. $\mathcal{N}(0,0.01)$. Observations are formed as $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}_r\boldsymbol{B}^\top + \boldsymbol{E}$. We solve the low-rank recovery problem using SCAD [37] and MCP [38]—as well as several existing baselines: nuclear norm and weighted nuclear norm. Table I reports the average reconstruction error across all datasets.

V. Concluison

In this paper, we have studied the problem of recovering low-rank matrices from noisy bilinear sketching measurements. We have proposed a novel estimator that minimizes a least-squares loss regularized by a nonconvex penalty to promote low-rank structure, and developed an efficient proximal gradient algorithm to solve the resulting nonconvex optimization problem. We have shown that the proposed method achieves the oracle convergence rate in the Frobenius norm under a minimal signal-strength condition. Numerical experiments on both synthetic and real-world datasets have validated the theoretical guarantees and demonstrated the practical effectiveness of the method.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on January 20,2026 at 15:39:03 UTC from IEEE Xplore. Restrictions apply.

References

[1] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[4] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[5] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa, "Compressive sensing for background subtraction," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 155–168.

[6] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[7] Z. Tian, Y. Tafesse, and B. M. Sadler, "Cyclic feature detection with sub-Nyquist sampling for wideband spectrum sensing," *IEEE J. Sel. Signal Process.*, vol. 6, no. 1, pp. 58–69, 2011.

[8] E. J. Candès *et al.*, "Compressive sampling," in *Proc. Int. Congr. Math.*, vol. 3. Madrid, Spain, 2006, pp. 1433–1452.

[9] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

[10] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York, USA: Springer-Verlag, 2010.

[11] S. Ganguli and H. Sompolinsky, "Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis," *Annu. Rev. Neurosci.*, vol. 35, no. 1, pp. 485–508, 2012.

[12] H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral, *Compressed sensing and its applications*. Cham, Switzerland: Birkhäuser, 2015, [online] Available: https://link.springer.com/book/ 10.1007/978-3-319-16042-9.

[13] M. Fornasier and H. Rauhut, "Compressive sensing." in *Handbook of Mathematical Methods in Imaging*. New York, USA, NY: Springer-Verlag, 2011, pp. 187–228.

[14] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[15] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053–4085, 2011.

[16] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[17] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Uniform uncertainty principle for Bernoulli and subgaussian ensembles," *Constructive Approx.*, vol. 28, pp. 277–289, 2008.

[18] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann, "Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling," *Constructive Approx.*, vol. 34, pp. 61–88, 2011.

[19] K. Zhong, P. Jain, and I. S. Dhillon, "Efficient Matrix Sensing Using Rank-1 Gaussian Measurements," in *Proc. 26th Int. Conf. Algorithmic Learn. Theory*, 2015, pp. 3–18.

[20] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and Stable Covariance Estimation From Quadratic Sampling via Convex Programming," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.

[21] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. D. Nowak, "Sketching sparse matrices, covariances, and graphs via tensor products," *IEEE Trans. Inf. Theory*, vol. 61, no. 3, pp. 1373–1388, 2015.

[22] H. Gui, J. Han, and Q. Gu, "Towards faster rates and oracle property for low-rank matrix estimation," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2016, pp. 2300–2309.

[23] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Stat.*, vol. 39, no. 5, pp. 2302 – 2923, 2011.

[24] A. C. Gilbert and K. Levchenko, "Compressing network graphs," in *Proc. LinkKDD Workshop 10th ACM Conf. KDD*, vol. 124, 2004.

[25] K. J. Ahn, S. Guha, and A. McGregor, "Graph sketches: sparsification, spanners, and subgraphs," in *Proc. ACM Symp. Princ. Database Syst.*, 2012, pp. 5–14.

[26] W. Wang and Z. Zhao, "Optimal compressive covariance sketching via rank-one sampling," in *Proc. Int. Conf. Sampling Theory Appl.*

[27] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase Retrieval using Alternating Minimization," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, pp. 2796–2804, 2013.

[28] D. D. Ariananda and G. Leus, "Compressive Wideband Power Spectrum Estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4775–4789, 2012.

[29] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," *Adv. Artif. Intell*, vol. 2009, 2009.

[30] W. Liu, C. Mu, R. Ji, S. Ma, J. Smith, and S.-F. Chang, "Low-Rank Similarity Metric Learning in High Dimensions," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2792–2799.

[31] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[32] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Annu. Asilomar Conf. Signal Syst. Comput.*, 1993, pp. 40–44.

[33] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pur. Appl Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[34] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *J. Roy. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[35] T. Zhang, "Analysis of Multi-stage Convex Relaxation for Sparse Regularization," *J. Mach. Learn. Res.*, vol. 11, no. 3, pp. 1081–1107, 2010.

[36] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing Sparsity by Reweighted $\ell1$ Minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877–905, 2008.

[37] J. Fan and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[38] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, pp. 894–942, 2010.

[39] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.

[40] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[41] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1665–1697, 2012.

[42] P.-L. Loh and M. J. Wainwright, "Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 559–616, Jan 2015.

[43] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *J. Mach. Learn. Res.*, vol. 11, pp. 2241–2259, 2010.

[44] Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *Ann. Stat.*, vol. 42, no. 6, p. 2164, 2014.

[45] J. Fan, H. Liu, Q. Sun, and T. Zhang, "I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error," *Ann. Stat.*, vol. 46, no. 2, p. 814, 2018.

[46] S. Negahban and M. J. Wainwright, "Estimation of (Near) Low-Rank Matrices with Noise and High-Dimensional Scaling," 2009, [online] Available: http://arxiv.org/PS_cache/arxiv/pdf/0912/0912.5100v1.pdf.

[47] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms.

[48] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, and Á. García-Martín, "Semantic-aware scene recognition," *Pattern Recognition*, vol. 102, p. 107256, 2020.

[49] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on January 20,2026 at 15:39:03 UTC from IEEE Xplore. Restrictions apply.