From Good Starts to Optimal Inference: Generalized Latent Factor Models with Missingness and Implicit Regularization

Chengzhu Huang Yuqi Gu

Department of Statistics, Columbia University

Abstract

Generalized latent factor models provide a flexible framework for analyzing high-dimensional non-Gaussian data, yet inference procedures under missingness remain largely underdeveloped. We study nonlinear latent factor models with exponential-family links and partially observed entries, and propose a unified computational and inferential pipeline that is both statistically optimal and algorithmically tractable. Our method proceeds in three stages: soft singular value thresholding for initialization, a one-step refinement that attains row-wise consistency for estimating the latent factors, and a vanilla gradient descent scheme that exploits implicit regularization to navigate the nonconvex likelihood landscape without explicit penalties. We show that gradient descent converges rapidly while preserving key identifiability structures under suitable initialization and learning rates. We establish a novel linear approximation of the gradient descent iterates and construct valid individual and simultaneous confidence bands for latent factors via Gaussian multiplier bootstrap. To our knowledge, this is the first provable framework that provides explicit guidance for initializing gradient descent and demonstrates that such properly initialized procedures can directly enable optimal statistical inference in nonlinear low-rank models with missing data. Extensive simulations confirm accurate estimation, reliable coverage, and robustness to missingness.

Keywords: Generalized Latent Factor Models; High-dimensional Inference; Nonconvex Optimization; Implicit Regularization; Nonlinear Matrix Completion; Gradient Descent.

1 Introduction

Latent factor models (Bartholomew et al., 2008) are widely used across many fields, including the social sciences, psychology, and recommendation systems. Because linear models are often inadequate for handling diverse discrete data types, generalized latent factor models (Bartholomew et al., 2008; Skrondal and Rabe-Hesketh, 2004; Huber et al., 2004; Kidzinski et al., 2022; Chen and Li, 2024) provide a more flexible framework that can accommodate binary, ordinal, and count data. More importantly, amid the rapid rise of large language models, the research community increasingly calls for trustworthy evaluation methods that can scale to a large number of models. Given the responses of different language models to a common set of questions, latent factor models offer a powerful tool for uncovering their underlying capabilities and distinctive strengths.

On the other hand, passive or incidental missingness is common in real-world applications. Traditional surveys inevitably contain missing values due to practical constraints. In the context of LLM evaluation, conducting a comprehensive assessment across all available problem sets is

This version: June 16, 2026.

both costly and time-consuming. As a result, it is often necessary to evaluate models using only a selected subset of questions.

To address missing data, a variety of methods along with theoretical guarantees have been developed for linear settings (Candes and Recht, 2012; Chen et al., 2019a). However, their extension to generalized latent factor models remains underdeveloped, making selective testing both necessary and prevalent in practice. More broadly, estimation and inference for nonlinear latent factor models under missingness are still not fully understood, let alone the extent to which missingness compromises the reliability of the resulting estimates. We focus on developing estimation and uncertainty quantification for generalized latent factor models with partially observed data in a unified, statistically optimal, and computationally tractable framework.

To formulate the problem, we consider models with exponential-family parameterizations, where the likelihood conditional on the latent factors and an observation index set $\Omega \subseteq [n] \times [p]$ is given by

$$l(\mathbf{R}|\boldsymbol{\zeta}, \mathbf{X}^*, \mathbf{Y}^*, \Omega) = \prod_{(i,j)\in\Omega} \exp\left[\gamma(R_{i,j}) + R_{i,j}M^*_{i,j} - \Psi(M^*_{i,j})\right] \tag{1}$$

with $\mathbf{M}^*_{i,j} \coloneqq \zeta^*_j + \mathbf{X}^{*\top}_i \mathbf{Y}^*_j$ for every $(i,j) \in \Omega$. Throughout, we adopt a uniform sampling model in which each entry belongs to Ω independently with probability π, which may *decay* with the dimensions n and p. As a concrete example, consider the logistic-link model for binary observations:

$$R_{i,j} \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\mathsf{Sigmoid}(\zeta^*_j + \mathbf{X}^{*\top}_i \mathbf{Y}^*_j)), \quad (i,j) \in \Omega,$$

where corresponds to the classic one-bit matrix completion problem (Davenport et al., 2014). As is standard in the latent variable literature, we impose a low-rank structure through the parameter matrix $\boldsymbol{\Theta}^*$. In contrast to the linear setting, however, this low-rank structure is observed through the nonlinear function Ψ, which constitutes the main challenge of the problem. In particular, this nonlinear formulation undermines two main strategies commonly used in prior work: (i) obtaining the maximizer of (1) in closed form, and (ii) conducting spectral analysis of the data matrix under a "low-rank signal plus noise" decomposition.

1.1 Non-convex Optimization Approaches and Local Landscape

A natural approach to maximizing the likelihood function (1) is to employ gradient-based methods for nonconvex optimization, which in turn requires a careful understanding of the local geometry of the objective, particularly its Hessian. In general, for regularized likelihoods under *full* observation, the landscape is often well behaved: within suitable ℓ_2 neighborhoods, the Hessian typically satisfies a form of positive definiteness and has a moderate condition number (Li et al., 2023; Ma et al., 2020):

$$\mathrm{Hess}(-\log l; \boldsymbol{\zeta}, \mathbf{X}, \mathbf{Y}) \overset{\text{up to log factors}}{\asymp} \mathrm{Hess}(-\log l^{\mathsf{pop}}; \boldsymbol{\zeta}^*, \mathbf{X}^*, \mathbf{Y}^*),$$

where l^{pop} denotes l's population counterpart. By contrast, missing data can severely distort the empirical Hessian as well as the optimization landscape, even in the linear setting, producing a large number of spurious suboptimal local minima. At first sight, this appears to undermine the viability of likelihood-based optimization. Fortunately, a series of prior works (Ma et al., 2018; Chen et al., 2019a, 2021) showed that, with proper initialization inside a *region of incoherence and contraction*, which requires row-consistency, gradient-based algorithms still enjoy Hessian properties analogous to those in the full-observation case. This naturally raises the following question in our setting:

> *In the presence of both nonlinearity and missingness, what characterizes the contractive region? More importantly, how can one reach this region in order to provide a warm start for gradient descent?*

As mentioned earlier, spectral estimators adopted in linear cases (Ma et al., 2018) no longer serve as reasonable starting points. On the other hand, when the link function Ψ is known *a priori*, several methods (Ma et al., 2020; Zhang et al., 2020) can consistently recover the latent structure, but only in the ℓ_2 sense. It remains unclear, however, whether these estimators are accurate enough to fall within the *region of incoherence and contraction*. These challenges therefore motivate us to develop a comprehensive warm-start scheme.

1.2 Linear Approximations and Downstream Inference for Gradient Descent

Despite the obstacle to proper initialization of gradient descent, one may expect to gain inferential benefits from hitting the stationary point within the *region of incoherence and contraction*. To see this, consider a *finite-dimensional* random smooth convex opitmization problem with an objective function L_n. Letting $\widehat{\boldsymbol{\theta}}_n$ be the stationary point of L_n, applying Taylor's expansion gives that

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* = \nabla^2 L_n(\widetilde{\boldsymbol{\theta}}_n)^{-1}\big[\nabla L_n(\widehat{\boldsymbol{\theta}}) - \nabla L_n(\boldsymbol{\theta}^*)\big] \approx -\nabla^2 \mathbb{E}[L_n](\boldsymbol{\theta}^*)^{-1}\nabla L_n(\boldsymbol{\theta}^*), \tag{2}$$

where $\widetilde{\boldsymbol{\theta}}_n$ resides on the line segment that connects $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$, and $\nabla^2 L_n(\widetilde{\boldsymbol{\theta}})^{-1}$ converges to its population counterpart. Then it suffices to examine the limiting distribution of $\nabla L_n(\boldsymbol{\theta}^*)$.

A seemingly straightforward implication is that, if an (approximate) stationary point is reached within the *region of incoherence and contraction*, we can similarly obtain a linear approximation $-\nabla^2 \mathbb{E}[L_n](\boldsymbol{\theta}^*)^{-1}\nabla L_n(\boldsymbol{\theta}^*)$ as a proxy of $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*$. [1] That said, extending this intuition from the fixed-dimensional regime to the high-dimensional setting is far from straightforward, especially in the presence of the rotational ambiguity inherent in factor parametrization. Several recent works (Wang, 2022; Li et al., 2023; Ouyang et al., 2024) have studied the statistical behavior of global minimizers of properly regularized maximum likelihood estimators. By contrast, our focus is algorithmic: we aim to establish a valid approximation for the outputs of computationally tractable procedures. This distinction is crucial because, in high-dimensional nonconvex problems, the existence of a well-behaved global minimizer does not by itself guarantee that a practical algorithm reaches such a point, nor does it directly yield the row-wise linear expansions needed for inference.

Beyond estimation itself, performing global comparisons across different profiles, items, or entries is both important and appealing in latent factor analysis, and lies at the heart of many ranking problems (Fan et al., 2025c). For instance, one may wish to compare individuals in terms of latent ability, items in terms of difficulty or discriminatory power, or more generally the relative magnitudes of latent components across heterogeneous groups. Such comparative questions often carry more direct substantive meaning than point estimation alone, as they reveal the underlying ordering, heterogeneity, and dominance relations encoded by the latent structure. As elaborated in Section 2.2 and Section 3, our main inferential results provide a solution by deriving an asymptotic representation for approximate stationary points and developing inference procedures for the aforementioned tasks.

1.3 Main Contributions

Building on the nonconvex optimization framework, this work makes substantial progress in understanding gradient descent dynamics in the presence of both nonlinearity and missingness, while addressing the accompanying initialization challenge through a carefully designed pipeline. Beyond estimation, we also develop flexible inferential procedures that support a variety of inference tasks.

[1] Here, $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ refer to some generic versions of latent factors or intercepts.

With the theoretical details are deferred to later sections, we summarize the main messages of our estimation and inference pipeline below:

1. We propose a nonlinear spectral decomposition procedure, incorperating the link function knowledge, to obtain Frobenius-norm-consistent estimators of $\boldsymbol{\zeta}^*$, $\mathbf{X}^*$, and $\mathbf{Y}^*$. Based upon this spectral estimator, we develop a one-step refinement procedure that upgrades Frobenius-norm consistency to $\ell_{2,\infty}$ consistency (row-wise consistency), thereby enabling the subsequent gradient analysis.

2. Given suitably initialized estimators with row-wise consistency, gradient descent enjoys a nearly linear convergence rate and eventually reaches a well-behaved stationary point. The required signal strength is primarily governed by the sampling rate and the singular value magnitude of $\mathbf{X}^*\mathbf{Y}^{*\top}$, both of which are shown to achieve the information-theoretic minimum order up to logarithmic factors.

3. We establish a novel linear approximation derived from the stationary conditions, and design a Gaussian multiplier bootstrap procedure that enables individual and simultaneous inference, as well as simultaneous missing-entry prediction.

4. To establish the results stated above, we need to carefully characterize the dynamics of gradient descent, via a celebrating mechanism called *implicit regularization*. Our analysis is underpinned by a leave-one-out argument conducted *throughout the entire three-step pipeline*, which obviates the need of extra data-splitting and constitutes additional technical contributions to account for the nonlinear observation mechanism.

2 Proposed Methods and Theoretical Results

The overall procedure consists of three steps, each of which plays a crucial role. The meta pipeline is presented as below.

Algorithmic Pipeline for Generalized Latent Factor Models

Input: Observation $\mathcal{P}_\Omega(\mathbf{R})$

`// Spectral Initialization.`

Apply the spectral algorithm (Algorithm 3) to obtain the initial estimates $(\widehat{\boldsymbol{\zeta}}^{\mathsf{spec}}, \widehat{\mathbf{X}}^{\mathsf{spec}}, \widehat{\mathbf{Y}}^{\mathsf{spec}})$;

`// One-step Refinement.`

Refine the spectral estimates by running the one-step refinement (Algorithm 4), yielding $(\widehat{\boldsymbol{\zeta}}^{\mathsf{os}}, \widehat{\mathbf{X}}^{\mathsf{os}}, \widehat{\mathbf{Y}}^{\mathsf{os}})$;

`// Vanilla Gradient Descent.`

Warm-start the gradient descent algorithm (Algorithm 2) using $(\widehat{\boldsymbol{\zeta}}^{\mathsf{os}}, \widehat{\mathbf{X}}^{\mathsf{os}}, \widehat{\mathbf{Y}}^{\mathsf{os}})$;

`// Inference.`

Perform inferential steps to quantify estimation uncertainty (see the methods in Section 3);

Output: Final estimates $(\widehat{\boldsymbol{\zeta}}^{t_0}, \widehat{\mathbf{X}}^{t_0}, \widehat{\mathbf{Y}}^{t_0})$ and associated uncertainty quantification.

For sake of clarity, however, we present the algorithmic analyses in an order reverse to that of the algorithm itself. We begin by analyzing the ingredients needed to warm-start gradient descent, together with the resulting estimation guarantees and linear approximation properties. We then turn to the earlier stages of the pipeline that lead to this warm start. Notably, the output of each

stage enjoys a progressively stronger form of consistency, ultimately meeting the requirements for initializing gradient descent.

2.1 Model Setup

We first take a moment to sort out the concrete setups, which accommondates a large class of continuous and discrete distributions.

Assumption 1 (Sampling scheme and noise assumptions)**.** *(a) Each entry is independently observed with probability* $\pi \in (0, 1]$.

(b) Each entry $R_{i,j}$*, whether it is observed or not, is independent, and is upper bounded by* B *with probability at least* $1 - O(d^{-c-10})$*. Their standard deviation are all upper bounded by* σ.

Further, we make assumptions on the smoothness of the link function in the exponential family. While the final error as well as the minimax lower bound are driven primarily by the second derivative of Ψ, the spectral initialization step as well as the injectivity of projection operator also require higher-order smoothness. Further discussions are provided in the Supplementary.

Assumption 2 (Link function conditions)**.** *Given the derivative* $\Psi'(x) =: \psi(x) : \mathbb{R} \supseteq \mathcal{D}_\psi \to \mathbb{R}$ *of* Ψ*, the function* $\psi(x)$ *obeys* $|\psi^{(k)}(x)| \le \bar{C}_\psi k! \bar{c}_\psi^k$ *for some constants* $\bar{c}_\psi$ *and* $\bar{C}_\psi$ *for every* $x \in \mathcal{D}_\psi$ *and* $k \in \mathbb{N}$*. Moreover, the second derivative of* ψ *is lower bound by* $\underline{c}_\psi$*. We write* $\kappa_\psi := \bar{c}_\psi / \underline{c}_\psi$ *throughout.*

Additionally, let $\sigma_1^*, \cdots, \sigma_r^*$ denote the top-r singular values of $\mathbf{X}^* \mathbf{Y}^{*\top}$, and define $\kappa := \sigma_1^*/\sigma_r^*$ and $d := n \vee p$. We introduce the following conditions on the low-rank signals.

Assumption 3 (Signal conditions)**.** *We make the following assumptions on the properties of* $\mathbf{X}^*$ *and* $\mathbf{Y}^*$*:*

(a) Signal Strength: *Assume that* $\sigma_r^* \gtrsim \sigma r \kappa^2 \kappa_\psi^2 \xi^4 \underline{c}_\psi^{-1} \sqrt{d/\pi}$ *and* $\pi \gtrsim \mu^3 r^2 \kappa^2 \kappa_\psi^2 \xi^4 / (n \wedge p)$*, where* $\xi := \left[c_\xi r \log d (\bar{c}_\psi \mu r / \sqrt{np} \vee 1) \right]^{\frac{r+1}{2}}$ *for some constant* c_ξ.

(b) Incoherence Degree: *Letting the top-*r *SVD of* $\mathbf{X}^* \mathbf{Y}^{*\top}$ *be* $(\mathbf{U}^*, \mathbf{\Lambda}^*, \mathbf{V}^*)$*, then the rows of* $\mathbf{U}^*$ *and* $\mathbf{V}^*$ *are upper bounded by* $\frac{n}{r} \|\mathbf{U}^*\|_{2,\infty}^2 \vee \frac{p}{r} \|\mathbf{V}^*\|_{2,\infty}^2 \le \mu$*. And there exists a scalar* ζ_0 *such that* $\|\boldsymbol{\zeta}^* - \zeta_0 \cdot \mathbf{1}_p\|_\infty \le \sqrt{\mu r \sigma_r^* / n^2}$*. Moreover, the relation to the entrywise upper bound* B *is given by* $\kappa^4 \kappa_\psi^4 \mu r^2 B \sqrt{\log d} \ll \sigma \sqrt{n \wedge p}$.

Remark 1. *Leaving* κ*,* κ_ψ*,* μ*, and* r *aside and taking the Gaussian case for example, our assumption on the signal strength matches the minimum requirement of this class of low-rank structure recovery problem Chen et al. (2019a). The conditions on the incoherence degree is widely adopted in the literature of low-rank structure recovery.*

Remark 2. *Our assumptions are designed to cover several challenging regimes that frequently arise in noisy and partially observed data. In particular, the theory allows for: (i) weak signal strength, with* $\sigma_r^* \gtrsim \sigma\sqrt{d}$ *up to logarithmic factors; (ii) severe missingness, with* π *as small as* $(n \wedge p)^{-1}$ *up to logarithmic factors; and (iii) weak local curvature of the link, with* $\underline{c}_\psi = o(1)$*. These regimes may overlap, and together they capture settings in which both statistical recovery and algorithmic stability are delicate.*

Identifiability Conditions. Before proceeding, we note that the parameters in our generalized latent factor model are not uniquely identifiable. To disentangle the intercepts from the latent factors, we treat $\boldsymbol{\zeta}^*$ as control feature sparsity and impose the following assumptions on the intercept and the factors:

Assumption 4 (Basic Identifiability)**.** *The factors* $\mathbf{X}^*$, $\mathbf{Y}^*$ *and the intercept vector* $\boldsymbol{\zeta}^*$ *obeys:*

$$\mathbf{1}^\top \mathbf{X}^* = \mathbf{0}, \tag{3}$$

$$\mathbf{X}^{*\top} \mathbf{X}^* = \rho \mathbf{Y}^{*\top} \mathbf{Y}^*. \tag{4}$$

Following the discussions in Bai and Li (2012); Fan et al. (2021), the parameter triple $(\boldsymbol{\zeta}^*, \mathbf{X}^*, \mathbf{Y}^*)$ is identifiable only up to an orthogonal rotation; that is, $(\mathbf{X}^*, \mathbf{Y}^*)$ and $(\mathbf{X}^*\mathbf{R}, \mathbf{Y}^*\mathbf{R})$ are equivalent for any $\mathbf{R} \in \mathcal{O}(r)$. In particular, the first condition in (3) centers the columns of $\mathbf{X}^*$, which renders $\boldsymbol{\zeta}^*$ identifiable. The second condition in (4) fixes the relative scaling between $\mathbf{X}^*$ and $\mathbf{Y}^*$. The scaling factor ρ is set *a priori* to accommodate Bayesian settings in which the latent factors naturally operate on different scales. For instance, if the entries of $\mathbf{X}^*$ and $\mathbf{Y}^*$ are independently drawn from their respective priors, then one typically has $\sigma_i(\mathbf{X}^*) \asymp \sqrt{n}$ and $\sigma_i(\mathbf{Y}^*) \asymp \sqrt{p}$. One may strenghten the identifiability slightly by removing the rotational ambiguity, as we shall elaborate in Section 2.2.

2.2 (Vanilla) Gradient Descent and Theoretical Guarantees

As all preceding steps are designed to guarantee the success of gradient descent, we begin by identifying the initialization conditions it requires and the consequences that follow. To this end, we consider the following objective function:

$$L(\boldsymbol{\zeta}, \mathbf{X}, \mathbf{Y}) = \sum_{(i,j)\in\Omega} -R_{i,j}(\zeta_j + \mathbf{x}_i^\top \mathbf{y}_j) + \Psi(\zeta_j + \mathbf{x}_i^\top \mathbf{y}_j) + c_{\mathsf{orth}} \frac{\sigma_1^*}{n} \left\| \mathbf{1}_n^\top \mathbf{X} \right\|_2^2, \tag{5}$$

where is essentially the negative log-likelihood function of the generalized latent factor model with a regularization term on the orthogonality between the all-one vector $\mathbf{1}_n$ and the column space of $\mathbf{X}$. The vanilla gradient descent algorithm is stated in Algorithm 2. The convergence rate is governed by the learning rate η, which can be chosen aggressively when the iterates remain in the RIC, as shortly elaborated in Section 2.3.

2 Vanilla Gradient Descent (GD)

Input: Initial estimates $\widehat{\boldsymbol{\zeta}}^0$, $\widehat{\mathbf{X}}^0$, $\widehat{\mathbf{Y}}^0$, learning rate η, number of iterations T

Gradient Descent Update: **for** $t = 1, 2, \ldots, t_0$ **do**

Update $(\boldsymbol{\zeta}, \mathbf{X}, \mathbf{Y})$ by

$$\begin{aligned}
\boldsymbol{\zeta}^t &= \boldsymbol{\zeta}^{t-1} - \eta \nabla_{\boldsymbol{\zeta}} L(\boldsymbol{\zeta}^{t-1}, \mathbf{X}^{t-1}, \mathbf{Y}^{t-1}), \\
\mathbf{X}^t &= \mathbf{X}^{t-1} - \eta \nabla_{\mathbf{X}} L(\boldsymbol{\zeta}^{t-1}, \mathbf{X}^{t-1}, \mathbf{Y}^{t-1}), \\
\mathbf{Y}^t &= \mathbf{Y}^{t-1} - \eta \nabla_{\mathbf{Y}} L(\boldsymbol{\zeta}^{t-1}, \mathbf{X}^{t-1}, \mathbf{Y}^{t-1}).
\end{aligned}$$

Rotational Adjustment: $\widehat{\mathbf{X}}^{\mathsf{rot}} = \mathbf{X}^{t_0}\widehat{\mathbf{R}}^\top$, $\widehat{\mathbf{Y}}^{\mathsf{rot}} = \mathbf{Y}^{t_0}\widehat{\mathbf{R}}^\top$, where $\widehat{\mathbf{R}}$ represents some appropriate rotation estimate.

Output: The factor estimates $\boldsymbol{\zeta}^{t_0}$, $\widehat{\mathbf{X}}^{\mathsf{rot}}$, $\widehat{\mathbf{Y}}^{\mathsf{rot}}$, and $\widehat{\boldsymbol{\Theta}} = \mathbf{1}_n {\boldsymbol{\zeta}^{t_0}}^\top + \widehat{\mathbf{X}}^{\mathsf{rot}} {\widehat{\mathbf{Y}}^{\mathsf{rot}}}^\top$.

Regularization Choice. We briefly discuss our choice of regularization. As noted in the introduction, row-wise consistency of the estimators helps rule out spurious landscape effects caused by missingness. A natural way to promote such behavior is to add a regularizer of the form $\|\mathbf{X}\|_{2,\infty}^2 + \|\mathbf{Y}\|_{2,\infty}^2$, though this comes at the price of additional computational complexity. Encouragingly, inspired by the seminal work of Ma et al. (2018), we show that even without such an explicit regularizer, the vanilla gradient descent iterates remain implicitly row-wise consistent, provided that the initialization already enjoys this property. On the other hand, prior works often imposed the penalty $\left\|\mathbf{X}^\top\mathbf{X} - \mathbf{Y}^\top\mathbf{Y}\right\|_F$ (assuming $\rho = 1$) to maintain balance between the scales of the left and right factors. Our analysis shows that this penalty is also unnecessary in our setting, as the same balancing effect can be achieved through a careful choice of learning rates. We defer a more detailed discussion of this implicit regularization phenomenon to Section 2.3.

We are now positioned to present our estimation error control for the gradient descent iterates.

Theorem 2.1. *Suppose that Assumptions 1, 2, 3, and 4 hold, and that* $c_{\mathsf{orth}} \gtrsim \kappa^2\kappa_\psi^2(\log d)^4$ *and* $\eta = \frac{c_\eta}{c_{\mathsf{orth}}\pi\sigma_1^*}$ *for some sufficiently small constant* c_η*. Then letting* $t_0 = \frac{c_t \log d}{\pi\sigma_r^*\eta}$ *for some sufficiently large* c_t*, with probability at least* $1 - (d^{-c})$*, the iterate* $(\boldsymbol{\zeta}^{t_0}, \mathbf{X}^{t_0}, \mathbf{Y}^{t_0})$ *satisfies*

$$\left\|\boldsymbol{\zeta}^{t_0} - \boldsymbol{\zeta}^*\right\|_2 \lesssim \frac{\sigma\sqrt{p\log d}}{\sqrt{\pi n}},\ \mathsf{dist}_{\mathrm{F}}(\mathbf{X}^{t_0}, \mathbf{X}^*) \lesssim \frac{\sigma\sqrt{nr\log d}}{\sqrt{\pi\sigma_r^*}}\rho^{\frac{1}{4}},\ \mathsf{dist}_{\mathrm{F}}(\mathbf{Y}^{t_0}, \mathbf{Y}^*) \lesssim \frac{\sigma\sqrt{pr\log d}}{\sqrt{\pi\sigma_r^*}}\rho^{-\frac{1}{4}}.$$

Remark 3. *The estimation error bounds are derived from a nearly linear convergence relation, taking* $\mathbf{X}^t$ *for example,* $\mathsf{dist}_F(\mathbf{X}^t, \mathbf{X}^*) \le (1 - \eta\underline{c}_\psi\pi\sigma_r^*)^t\mathsf{dist}_F(\mathbf{X}^0, \mathbf{X}^*) + \frac{\sigma\sqrt{nr\log d}}{\sqrt{\pi\sigma_r^*}}$*. Similar relations holds in the sense of* $\ell_{2,\infty}$ *as well as for* $\mathbf{Y}^*$ *and* $\boldsymbol{\zeta}^*$*.*

Remark 4. *We remark on the range of learning rates allowed by our theory. In line with the implicit-regularization phenomenon identified by Ma et al. (2018), our analysis permits learning rates that are nearly as large as the stability limit for gradient descent, up to logarithmic factors.*

More importantly, beyond the estimation error control described above, we provide a more refined characterization by quantifying the error in the linear approximation of the gradient descent iterates.

Theorem 2.2. *Instate the assumptions in Theorem 2.1. Then with probability at least* $1 - O(d^{-c})$*, the proximity of the final output to the gradient forms at the ture parameter is characterized by*

$$\max_{j\in[n]}\left|\left(\zeta_j^{t_0} - \zeta_j^*\right) + \Big[\sum_{i\in[n],(i,j)\in\Omega}\psi'(\zeta_j^* + \mathbf{X}_i^{*\top}\mathbf{Y}_j^*)\Big]^{-1}(\mathcal{P}_{\mathbf{\Omega}}(\mathbf{E}))^\top\mathbf{1}_n\right| \lesssim \frac{\sigma}{\sqrt{\pi n\log d}},$$

$$\max_{i\in[n]}\left\|\left(\mathbf{X}_{i,\cdot}^{t_0}\mathsf{sgn}(\mathbf{X}^{t_0\top}\mathbf{X}^*) - \mathbf{X}_{i,\cdot}^*\right) + (\nabla_{\mathbf{X}}L^\circ)_{i,\cdot}\mathbf{Y}^*\left[\mathbf{Y}^{*\top}\mathsf{diag}\big(\psi'(\zeta_j^* + \mathbf{X}_i^{*\top}\mathbf{Y}_j^*)\big)_{j\in[p]}\mathbf{Y}^*\right]^{-1}\right\|_2 \lesssim \frac{\sigma\sqrt{r}\rho^{\frac{1}{4}}}{\sqrt{\pi\sigma_r^*\log d}},$$

$$\max_{j\in[p]}\left\|\left(\mathbf{Y}_{j,\cdot}^{t_0}\mathsf{sgn}(\mathbf{Y}^{t_0\top}\mathbf{Y}^*) - \mathbf{Y}_{j,\cdot}^*\right) + (\nabla_{\widetilde{\mathbf{Y}}}L^\circ)_{\cdot,j}^\top\mathbf{X}^*\left[\mathbf{X}^{*\top}\mathsf{diag}\big(\psi'(\zeta_j^* + \mathbf{X}_i^{*\top}\mathbf{Y}_j^*)\big)_{i\in[n]}\mathbf{X}^*\right]^{-1}\right\|_2 \lesssim \frac{\sigma\sqrt{r}\rho^{-\frac{1}{4}}}{\sqrt{\pi\sigma_r^*\log d}},$$

where $\nabla_{\mathbf{X}}L^\circ$ *and* $\nabla_{\widetilde{\mathbf{Y}}}L^\circ$ *are defined as* $(\mathcal{P}_{\mathbf{\Omega}}(\mathbf{E}))\mathbf{Y}^*$ *and* $(\mathcal{P}_{\mathbf{\Omega}}(\mathbf{E}))^\top\widetilde{\mathbf{X}}^*$*, resp., with* $\widetilde{\mathbf{X}}^* \coloneqq (\sqrt{\frac{\sigma_r^*}{n}}\mathbf{1}, \mathbf{X}^*)$*.*

To illustrate the scope of our theory, we highlight the following implications.

- Sample Complexity The strength of our theory is reflected, in part, by its allowance for a vanishing sampling rate down to the information-theoretic threshold $\frac{1}{n\wedge p}$. This stands in contrast to prior work on nonlinear low-rank models Ouyang et al. (2024); Ma et al. (2020);

Li et al. (2023); Su and Wang (2025), which either does not allow missingness or only permits sampling rates of constant order. The latter regime is substantially more tractable than the one considered here, since extensive missingness destroys the local ℓ_2 convexity of the landscape and thus rely on the analysis in Section 2.3.

- No need for debiasing. The linear approximation is readily seen to be mean zero. In contrast to convex programming approaches (Chen et al., 2019b) for the linear matrix completion, our method does not require an additional debiasing step, since the underlying optimization procedure avoids bias-inducing regularization.

Rotation Adjustment. Although the rotation aligning $(\mathbf{X}^{t_0}, \mathbf{Y}^{t_0})$ with $(\mathbf{X}^*, \mathbf{Y}^*)$ is generally not identifiable, it can be fixed under additional population-level structure. One possibility is to assume that the signal matrix $\mathbf{X}^*\mathbf{Y}^{*\top}$ has distinct singular values with sufficiently large eigengaps, in which case the singular subspaces determine a canonical orientation up to column-wise sign changes. Another possibility is to exploit distributional structure in the loadings: if the entries of $\mathbf{Y}^*$, interpreted as item discriminations, are drawn from a non-Gaussian distribution with sufficiently pronounced fourth-order structure, then rotational invariance is broken and an interpretable criterion such as varimax (Rohe and Zeng, 2023) can be used to select a canonical rotation. The statistical guaranties resulted from these conditions are further discussed in the Supplementary.

Minimax Lower Bounds (Optimality). Following standard arguments in information lower bounds, we have the following minimax rates of our models, which are matched by the bounds in Theorem 2.1.

Theorem 2.3. *Consider the Bernoulli distribution family with the following parameter space:*

$$\begin{aligned}\boldsymbol{\Theta} \coloneqq \Big\{(\boldsymbol{\zeta}^*, \mathbf{X}^*, \mathbf{Y}^*) : \|\boldsymbol{\zeta}^* - c_\zeta \mathbf{1}\|_\infty \le c_1, \|\mathbf{X}^*\|_{2,\infty} \le c_2(\sigma_r^*/n)^{\frac{1}{2}}, \|\mathbf{Y}^*\|_{2,\infty} \le c_2(\sigma_r^*/p)^{\frac{1}{2}},\\ \operatorname{rank}(\mathbf{X}^*\mathbf{Y}^{*\top}) = r, \sigma_r(\mathbf{X}^*\mathbf{Y}^{*\top}) \in [0.9\sigma_r^*, 1.1\sigma_r^*]\Big\}\end{aligned}$$

where the rank r and c_i are constants. Then there exists some universal constant c_{lb} such that

$$\inf_{\widehat{\mathbf{X}}} \sup_{(\boldsymbol{\zeta}^*, \mathbf{X}^*, \mathbf{Y}^*) \in \boldsymbol{\Theta}} \mathbb{E}\big[\min_{\mathbf{R} \in \mathcal{O}(r)} \big\|\widehat{\mathbf{X}}\mathbf{R} - \mathbf{X}^*\big\|\big] \vee \inf_{\widehat{\mathbf{Y}}} \sup_{(\boldsymbol{\zeta}^*, \mathbf{X}^*, \mathbf{Y}^*) \in \boldsymbol{\Theta}} \mathbb{E}\big[\min_{\mathbf{R} \in \mathcal{O}(r)} \big\|\widehat{\mathbf{Y}}\mathbf{R} - \mathbf{Y}^*\big\|\big] \ge c_{\mathsf{lb}} \frac{\sqrt{p}}{\underline{c}_\psi \sqrt{\pi \sigma_r^*}}.$$

2.3 Technical Challenges

Several highlights of our technical contributions of the gradient descent analysis are in order. We begin by characterizing the optimization landscape, and then discuss the key technical ingredients underlying the linear approximations.

2.3.1 Underlying Mechanism: Implicit Regularization

We now shed light on the behavior of gradient descent iterates and on why a warm start is necessary. As a point of reference, consider the linear setting of low-rank matrix completion. The seminal work (Ma et al., 2018) offered a new perspective through the lens of implicit regularization, showing that gradient descent can automatically constrain its trajectory to remain within the *region of incoherence and contraction (RIC)*. This property makes it possible to obtain precise control of

the Hessian along tangent directions at any point in the RIC. In this paper, we show that this implicit regularization phenomenon extends to generalized latent factor models with missingness. Moreover, in our analysis, implicit regularization manifests itself in two distinct senses: (i) locational regularization within the *RIC*, and (ii) approximate balancedness between $\mathbf{X}^*$ and $\mathbf{Y}^*$.

Implicit Restriction to the RIC. We start with the characterization of the *region of incoherence and contraction*, assuming $\mu = O(1)$. The region consists of parameters that are close to $(\boldsymbol{\zeta}^*, \mathbf{X}^*, \mathbf{Y}^*)$ in the sense of spectral norm and ℓ_2 norm:

$$\mathrm{RIC} := \Bigg\{ (\boldsymbol{\zeta}, \mathbf{X}, \mathbf{Y}) : \left\|\mathbf{1}^\top \mathbf{X}\right\|_2 \ll \sqrt{\xi n \sigma_{\min}}, \\ \max\left\{ \frac{\|\mathbf{X}-\mathbf{X}^*\|_{2,\infty}}{\|\mathbf{X}^*\|_{2,\infty}}, \frac{\|\mathbf{Y}-\mathbf{Y}^*\|_{2,\infty}}{\|\mathbf{Y}^*\|_{2,\infty}}, \|\boldsymbol{\zeta}-\boldsymbol{\zeta}^*\|_\infty \right\} \ll \frac{1}{\sqrt{\kappa^3 \mu r \log^2 d}} \Bigg\}$$

We assert the restricted strong convexity and smoothness for genearlized latent factor models with missing data in the *RIC*, whose informal statement is as follows:

Informal Theorem 2.4. *Suppose the sampling rate satisfies* $\pi \gtrsim \frac{\mathrm{polylog}(d)}{\sqrt{n \wedge p}}$. *Then, with overwhelming probability, the* curvature *of the reparameterized objective* $L\big(\sqrt{\sigma_1^*/n}\,\boldsymbol{\zeta}', \mathbf{X}, \mathbf{Y}\big)$, *where* L *is defined in* (5), *is bounded from below and above by* $\underline{c}_\psi \pi \sigma_r^*$ *and* $\bar{c}_\psi \pi \sigma_1^* \sqrt{\log d}$, *respectively, along directions that approximately lie in the tangent space at* $(\boldsymbol{\zeta}', \mathbf{X}, \mathbf{Y})$.

The preceding result implies linear convergence, with a contraction factor proportional to the learning rate, as long as the iterates remain inside the *RIC*. This invariance of the trajectory is not a consequence of local strong convexity alone; it depends on finer structural properties of the loss and the factorized parametrization. Whereas some prior works enforce such stability through explicit regularization (Ma et al., 2020), we show, following the implicit-regularization perspective of Ma et al. (2018); Chen et al. (2019a, 2021), that properly initialized vanilla gradient descent stays in the *RIC* automatically.

Informal Theorem 2.5. *Suppose the gradient descent algorithm is initialized at a point* $(\boldsymbol{\zeta}^0, \mathbf{X}^0, \mathbf{Y}^0)$ *that lies in the* RIC *with high probability. Then, with high probability, all iterates remain in the* RIC *for every* $t \le T$.

Note that this implicit regularization analysis — particularly the control of $\mathsf{dist}_{2,\infty}(\mathbf{X}^t, \mathbf{X}^*)$ and $\mathsf{dist}_{2,\infty}(\mathbf{Y}^t, \mathbf{Y}^*)$ — relies crucially on a leave-one-out argument, which decouples the dependence between the iterates $(\boldsymbol{\zeta}^t, \mathbf{X}^t, \mathbf{Y}^t)$ and any fixed row- or column-wise noise vector $\big[\widetilde{\mathbf{R}} - \pi\psi(\mathbf{1}\boldsymbol{\zeta}^{*\top} + \mathbf{X}^*\mathbf{Y}^{*\top})\big]_{i,\cdot}$ or $\big[\widetilde{\mathbf{R}} - \pi\psi(\mathbf{1}\boldsymbol{\zeta}^{*\top} + \mathbf{X}^*\mathbf{Y}^{*\top})\big]_{\cdot,j}$. Carrying out such an argument for the gradient descent stage, however, requires leave-one-out analysis for all stages preceding gradient descent, which poses additional technical challenges. We defer a detailed discussion of this point to Section 4.

Implicit Balancedness We also remark that throughout the iterates, the gradient descent algorithm is able to maintain the balancedness conditions appearing in the identifiability condition (3), without explicitly imposing regularization such as $\left\|\mathbf{X}^\top\mathbf{X} - \mathbf{Y}^\top\mathbf{Y} \cdot \rho\right\|_F$.

Informal Theorem 2.6. *Suppose that* ${\mathbf{X}^0}^\top\mathbf{X}^0 = {\mathbf{Y}^0}^\top\mathbf{Y}^0$ *and* $\mathbf{1}_n^\top\mathbf{X}^0 = \mathbf{0}$. *The gradient descent algorithm with an appropriately chosen learning rate will returns a sequence of iterates such that* $\left\|\rho^{-\frac{1}{2}}{\mathbf{X}^T}^\top\mathbf{X}^T - {\mathbf{Y}^T}^\top\mathbf{Y}^T \cdot \rho^{\frac{1}{2}}\right\|$ *is negligible with high probability.*

These forms of implicit regularization are noteworthy for two reasons. First, unlike Tu et al. (2016); Zheng and Lafferty (2016); Park et al. (2018), we do not impose any explicit regularization term to guarantee them. Second, they resonate with the discussion in Soltanolkotabi et al. (2023) for matrix sensing, but here they arise in a subtler notion of balancedness that is essential for the downstream inferential procedures.

2.3.2 Key Ingradients of Linear Approximations

Given that the gradient descent iterates eventually converge to a stationary point within the RIC, we now briefly outline how the linear approximation in Part 2 of Theorem 2.1 is derived from the stationary condition. This derivation constitutes the main technical contribution of our uncertainty quantification analysis.

Let $\widehat{\boldsymbol{\theta}}^{t_0}$ denote the concatenation of $(\widehat{\boldsymbol{\zeta}}^{t_0}, \widehat{\mathbf{X}}^{t_0}, \widehat{\mathbf{Y}}^{t_0})$. At a high level, the desired linear approximation for an approximate stationary point simply takes the form (2). However, the objective function under consideration is nonconvex and, even locally, lacks strong convexity in all directions. This degeneracy arises from the factorized parametrization: the latent factors are invariant under mutual rescalings and rotations, which makes the Hessian matrix singular.

To address this issue, we carefully decompose the parameter space into mutual scaling directions, rotational directions, and valid error directions. Since the leading estimation error lies primarily along the valid error directions, we use the implicit regularization characterization established above to control the error components along the remaining directions. Moreover, the low-rank structure consists not only of two low-dimensional factors, but also of an intercept component, which is not necessarily orthogonal to the column space of $\mathbf{Y}^*$. This creates nontrivial dependence among the resulting estimators. We overcome this difficulty by adopting the augmented reparametrization $\widetilde{\mathbf{X}} = \big(\sqrt{\frac{\sigma_r^*}{n}}\mathbf{1}, \mathbf{X}^*\big)$, $\widetilde{\mathbf{Y}} = \big(\sqrt{\frac{n}{\sigma_r^*}}\boldsymbol{\zeta}^*, \mathbf{Y}^*\big)$, throughout the analysis, which enables a clean characterization of the joint distribution.

3 Individual and Simultaneous Uncertainty Quantification

So far we have presented a linear approximation consistency in Theorem 2.1. This places us in a position to introduce our methodological framework for uncertainty quantification, enabled by these linear approximations. We shall start from individual inference, as a direct consequence of asymptotic normality. Going beyond, we extend the scenario to the simultaneous inference, leveraging the Gaussian multiplier bootstrap technique.

Individual Inference for single entry/row. Given the zero mean nature of the gradient terms appearing in Theorem 2.2, asymptotic normality typically follows under mild condition on the noise, as formally stated in the following.

Theorem 3.1. *Instate the assumptions in Theorem 2.1. Then the rows of $\boldsymbol{\zeta}^{t_0}$, $\mathbf{X}^{t_0}$, and $\mathbf{Y}^{t_0}$ obey*

$$\sup_{x\in\mathbb{R}}\Big|\mathbb{P}\Big[\big(\nabla^2_{\boldsymbol{\zeta},j_0}L(\boldsymbol{\zeta}^*,\mathbf{X}^*,\mathbf{Y}^*)\big)^{\frac{1}{2}}\big(\zeta_{j_0}^{t_0}-\zeta_{j_0}^*\big)\le x\Big]-\mathbb{P}_{z\sim\mathcal{N}(0,1)}\big[z\le x\big]\Big| = o(1),$$

$$\sup_{\text{convex set } A\subseteq\mathbb{R}^r}\Big|\mathbb{P}\Big[\big(\nabla^2_{\mathbf{X},i_0}L(\boldsymbol{\zeta}^*,\mathbf{X}^*,\mathbf{Y}^*)\big)^{\frac{1}{2}}\big(\mathbf{X}_{i_0}^{t_0}-\mathbf{X}_{i_0}^*\big)\in A\Big]-\mathbb{P}_{z\sim\mathcal{N}(0,\mathbf{I}_r)}\big[z\in A\big]\Big| = o(1),$$

$$\sup_{\text{convex set } A\subseteq\mathbb{R}^r}\Big|\mathbb{P}\Big[\big(\nabla^2_{\mathbf{Y},j_0}L(\boldsymbol{\zeta}^*,\mathbf{X}^*,\mathbf{Y}^*)\big)^{\frac{1}{2}}\big(\mathbf{Y}_{j}^{t_0}-\mathbf{Y}_{j}^*\big)\in A\Big]-\mathbb{P}_{z\sim\mathcal{N}(0,\mathbf{I}_r)}\big[z\in A\big]\Big| = o(1),$$

where, for arbitrary $i_0\in[n]$, $j_0\in[p]$ and a paremeter tuple $(\boldsymbol{\zeta},\mathbf{X},\mathbf{Y})$, the shorthand notations $\nabla^2_{\boldsymbol{\zeta}_j}L(\boldsymbol{\zeta},\mathbf{X},\mathbf{Y}), \nabla^2_{\mathbf{X}_i}L(\zeta,\mathbf{X},\mathbf{Y})$, $\nabla^2_{\mathbf{Y}_j}(\zeta,\mathbf{X},\mathbf{Y})$ denote that

$$\nabla^2_{\mathbf{X},i_0}L(\boldsymbol{\zeta},\mathbf{X},\mathbf{Y}) \coloneqq \sum_{j\in[p],(i_0,j)\in\Omega}\psi'(\zeta_j+\mathbf{X}_{i_0}{}^\top\mathbf{Y}_j)\mathbf{Y}_j{}^\top\mathbf{Y}_j,$$

$$\nabla^2_{\widetilde{\mathbf{Y}}_{j_0}}L(\boldsymbol{\zeta},\mathbf{X},\mathbf{Y}) \coloneqq \sum_{i\in[n],(i,j_0)\in\Omega}\psi'(\zeta_{j_0}+\mathbf{X}_i{}^\top\mathbf{Y}_{j_0})\widetilde{\mathbf{X}}_i^\top\widetilde{\mathbf{X}}_i.$$

A natural way to estimate the covariance matrices $\nabla^2_{*,*}L(\boldsymbol{\zeta}^*, \mathbf{X}^*, \mathbf{Y}^*)$ appearing in these asymptotic distributions is to plug the gradient descent estimate $(\widehat{\boldsymbol{\zeta}}^{t_0}, \widehat{\mathbf{X}}^{t_0}, \widehat{\mathbf{Y}}^{t_0})$ into the corresponding covariance formulas. Formally, for the corresponding plug-in confidence band, we establish the following coverage guarantees.

Theorem 3.2. *Instate the assumptions in Theorem 2.1. Then the coverage probability can be controlled by*

$$\sup_{0<\alpha<1}\Big|\mathbb{P}\Big[\zeta^*_{j_0} \in \mathcal{I}_{\boldsymbol{\zeta},j_0}\Big] - (1-\alpha)\Big| = o(1), \quad \sup_{0<\alpha<1}\Big|\mathbb{P}\Big[\mathbf{X}^*_{i_0} \in \mathcal{I}_{\mathbf{X},i_0}\Big] - (1-\alpha)\Big| = o(1),$$
$$\sup_{0<\alpha<1}\Big|\mathbb{P}\Big[\mathbf{Y}^*_{j_0} \in \mathcal{I}_{\mathbf{Y},j_0}\Big] - (1-\alpha)\Big| = o(1),$$

where the confidence intervals/bands are defined as

$$\mathcal{I}_{\boldsymbol{\zeta},j} \coloneqq \Big\{x \in \mathbb{R} : \big[(\nabla^2_{\widetilde{\mathbf{Y}}_{j_0}} L(\widehat{\boldsymbol{\zeta}}^{t_0}, \widehat{\mathbf{X}}^{t_0}, \widehat{\mathbf{Y}}^{t_0})^{-1}\big)\big]^{-\frac{1}{2}}_{1,1}(x - \widehat{\zeta}^{t_0}_{j_0}) \le z_{\frac{\alpha}{2}}\Big\}$$
$$\mathcal{I}_{\mathbf{X},i} \coloneqq \Big\{\mathbf{x} \in \mathbb{R}^r : \Big\|\big(\nabla^2_{\mathbf{X},i_0} L(\widehat{\boldsymbol{\zeta}}^{t_0}, \widehat{\mathbf{X}}^{t_0}, \widehat{\mathbf{Y}}^{t_0})\big)^{\frac{1}{2}}\big(\mathbf{x} - \mathbf{X}_i^{(t_0)}\big)\Big\|_2^2 \le q_{\chi,r,\alpha}\Big\},$$
$$\mathcal{I}_{\mathbf{Y},j} \coloneqq \Big\{\mathbf{y} \in \mathbb{R}^r : \Big\|\big[(\nabla^2_{\widetilde{\mathbf{Y}}_{j_0}} L(\widehat{\boldsymbol{\zeta}}^{t_0}, \widehat{\mathbf{X}}^{t_0}, \widehat{\mathbf{Y}}^{t_0})^{-1}\big)\big]^{-\frac{1}{2}}_{2:K+1,2:K+1}\big(\mathbf{y} - \mathbf{Y}_j^{(t_0)}\big)\Big\|_2^2 \le q_{\chi,r,\alpha}\Big\}.$$

Here, $z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$-quantile of the standard normal distribution, and $q_{\chi,r,\alpha}$ is the α-quantile of the χ^2 distribution with r degrees of freedom.

In assessing the optimality of our inference procedure, we observe that the inverses of the normalizing terms match the Cramèr–Rao lower bounds for estimating a single entry or row of $\boldsymbol{\zeta}^*$, $\mathbf{X}^*$, or $\mathbf{Y}^*$, when the remaining parameters are treated as known oracle quantities.

Remark 5. *In assessing the optimality of our inference procedure, we observe that the inverses of the normalizing terms match the Cramèr–Rao lower bounds for estimating a single entry or row of $\boldsymbol{\zeta}^*$, $\mathbf{X}^*$, or $\mathbf{Y}^*$, when the remaining parameters are treated as known oracle quantities. Formally, suppose that we have access to the intercept $\boldsymbol{\zeta}^*$ as well as the right factor $\mathbf{Y}^*$. Then for any unbiased estimator $\widehat{\mathbf{u}}$ for $\mathbf{X}^*_{i_0}$, we have*

$$\mathrm{Cov}(\widehat{\mathbf{u}} \mid \Omega) \succeq \sum_{j\in[p],(i_0,j)\in\Omega} \psi'(\zeta^*_j + \mathbf{X}^{*\top}_{i_0}\mathbf{Y}^*_j)\mathbf{Y}^{*\top}_j\mathbf{Y}^*_j \coloneqq \mathrm{CRLB}(\mathbf{X}^*_{i'} \mid \Omega).$$

This lower bound coincides with the limiting covariance of our estimator in Theorem 3.1, thereby showing its oracle efficiency. The same argument applies to inference for $\boldsymbol{\zeta}^*$ and $\mathbf{Y}^*$.

Simultaneous Confidence Bands via Gaussian Multiplier Bootstrap Despite quantifying the uncertainty for fixed number of parameters, there are many scenarios that require simultaneous evaluation for a growing number of paremeters. Thanks to the linear approximations of the estimators that uniformly hold in the row-wise sense, we are able to simulate distributions of certain statistics (e.g. the maximum statistic) regarding many parameters via the Gaussian multiplier bootstrap.

We start with the general framework that takes into account both factor-wise and entrywise-wise inference. Precisely, consider the statistics

$$\mathcal{T} \coloneqq \max_{q\in[m]}\Big\{\sum_{(i_1,k_1)\in\mathcal{S}_{q,1}\subseteq[n]\times[r]} a_{i_1,k_1}\big(\widehat{X}^{t_0}_{i_1,k_1} - X^*_{i_1,k_1}\big)$$

$$+ \sum_{(i_2,k_2)\in\mathcal{S}_{q,2}\subseteq[p]\times[r]} b_{i_2,k_2}\big(\widehat{\widetilde{Y}}^{t_0}_{i_2,k_2} - \widetilde{Y}^{*}_{i_2,k_2}\big) + \sum_{(k,l)\in\mathcal{S}_{q,3}\subseteq[n]\times[p]} c_{k,l}\big(\widehat{M}^{t_0}_{k,l} - M^{*}_{k,l}\big)\Big\}$$

for the index sets $\{(\mathcal{S}_{q,1},\mathcal{S}_{q,2},\mathcal{S}_{q,3})\}_{q\in[m]}$. As need in most application scenerios and assumed in our theory, only few elements of $\{a_{i_1,k_2}\}$, $\{b_{i_2,k_2}\}$, and $\{c_{k,l}\}$ are non-zero while we allow for a large m. In order to mimic the distribution of $\mathcal{T}$,

1. Fix a sufficiently large bootstrap size B. For each $b\in[B]$, generate independent standard Gaussian multipliers
$$\{\xi^{[b]}_{i,\mathbf{X}} : i\in[n]\}, \qquad \{\xi^{[b]}_{j,\mathbf{Y}} : j\in[p]\}.$$

2. For each bootstrap replication $b\in[B]$, construct the multiplier-bootstrap counterpart of the gradient terms
$$\nabla_{\mathbf{X}_i} L^{[b]}(\widehat{\boldsymbol{\zeta}}^{t_0},\widehat{\mathbf{X}}^{t_0},\widehat{\mathbf{Y}}^{t_0}) := \big[(\psi(\mathbf{1}\widehat{\boldsymbol{\zeta}}^{T\top} + \widehat{\mathbf{X}}^{t_0}\widehat{\mathbf{Y}}^{t_0\top}) - \mathbf{R})\circ\boldsymbol{\Omega}\big]_{i,\cdot}\mathsf{diag}(\xi^{[b]}_{j,\mathbf{Y}})_{j\in[p]}\widehat{\mathbf{Y}}^{t_0}, \qquad i\in[n],$$
$$\nabla_{\widetilde{\mathbf{Y}}_j} L^{[b]}(\widehat{\boldsymbol{\zeta}}^{t_0},\widehat{\mathbf{X}}^{t_0},\widehat{\mathbf{Y}}^{t_0}) := \big[(\psi(\mathbf{1}\widehat{\boldsymbol{\zeta}}^{T\top} + \widehat{\mathbf{X}}^{t_0}\widehat{\mathbf{Y}}^{t_0\top}) - \mathbf{R})\circ\boldsymbol{\Omega}\big]^{\top}_{\cdot,j}\mathsf{diag}(\xi^{[b]}_{i,\mathbf{X}})_{i\in[n]}\widehat{\mathbf{X}}^{t_0}, \qquad j\in[p].$$

   Based upon these components, we introduce the linearized error terms
$$(\Delta_{\mathbf{X}})_{i,\cdot} = \Delta_{\mathbf{X}_i} := -\nabla_{\mathbf{X}_i} L^{[b]}(\widehat{\boldsymbol{\zeta}}^{t_0},\widehat{\mathbf{X}}^{t_0},\widehat{\mathbf{Y}}^{t_0})\big(\nabla^2_{\mathbf{X}_i} L^{[b]}(\widehat{\boldsymbol{\zeta}}^{t_0},\widehat{\mathbf{X}}^{t_0},\widehat{\mathbf{Y}}^{t_0})\big)^{-1},$$
$$(\Delta_{\widetilde{\mathbf{Y}}})_{j,\cdot} = \Delta_{\widetilde{\mathbf{Y}}_j} := -\nabla_{\widetilde{\mathbf{Y}}_j} L^{[b]}(\widehat{\boldsymbol{\zeta}}^{t_0},\widehat{\mathbf{X}}^{t_0},\widehat{\mathbf{Y}}^{t_0})\big(\nabla^2_{\widetilde{\mathbf{Y}}_j} L^{[b]}(\widehat{\boldsymbol{\zeta}}^{t_0},\widehat{\mathbf{X}}^{t_0},\widehat{\mathbf{Y}}^{t_0})\big)^{-1}.$$

3. We approximate the distribution of $\mathcal{T}$ by the empirical distribution, over $b\in[B]$, of
$$\mathcal{T}^{[b]} := \max_{q\in[m]}\Big\{ \sum_{(i_1,k_1)\in\mathcal{S}_{q,1}\subseteq[n]\times[r]} \widehat{a}_{i_1,k_1}(\Delta^{[b]}_{\mathbf{X}})_{i_1,k_1} + \sum_{(i_2,k_2)\in\mathcal{S}_{q,2}\subseteq[p]\times[r]} \widehat{b}_{i_2,k_2}(\Delta^{[b]}_{\mathbf{Y}})_{i_2,k_2}$$
$$+ \sum_{(k,l)\in\mathcal{S}_{q,3}\subseteq[n]\times[p]} \widehat{c}_{k,l}\Big[(\Delta^{[b]}_{\mathbf{X}})_{k,\cdot}\widehat{\mathbf{Y}}^{t_0\top}_{l,\cdot} + (\mathbf{1}_n,\widehat{\mathbf{X}}^{t_0})_{k,\cdot}(\Delta^{[b]}_{\widetilde{\mathbf{Y}}})^{\top}_{l,\cdot}\Big]\Big\}.$$

The validity of this procedure can be justified by the results on high-dimensional central limit theorems (Chernozhukov et al., 2013; Chernozhuokov et al., 2022; Chernozhukov et al., 2023a), as presented in the following theorem.

Theorem 3.3. *Instate the assumptions in Theorem 2.1. Then, under some appropriate regularity conditions, the empirical distributions of the statistics introduced in Step 3 obey with probability* $1-o(1)$ *that*

$$\max_{x\in\mathbb{R}^+}\big|\mathbb{P}\big[\max_{i_0\in S_X}\big|\widehat{X}^{t_0}_{i_0,r_1} - X^{*}_{i_0,r_1}\big|\le x\big] - \mathbb{P}\big[L^{[b]}_{\mathbf{X},S_X}\le x \mid \mathbf{R},\Omega\big]\big| = o(1),$$
$$\max_{x\in\mathbb{R}^+}\big|\mathbb{P}\big[\max_{(i_1,j_1)\in S}\big|\widehat{\zeta}^{t_0}_{j_1} + \widehat{\mathbf{X}}^{t_0\top}_{i_1}\widehat{\mathbf{Y}}^{t_0}_{j_1} - \zeta^{*}_{j_1} - \mathbf{X}^{*\top}_{i_1}\mathbf{Y}^{*}_{j_1}\big|\le x\big] - \mathbb{P}\big[L^{[b]}_{S}\le x \mid \mathbf{R},\Omega\big]\big| = o(1).$$

Ranking Intervals for Factor Entries An important downstream task after recovering the latent structure is to rank individuals according to their abilities. In psychometrics, for example, one is naturally interested in comparing the latent abilities of participants, while in the evaluation of large language models, a key goal is to assess their subject-matter expertise. These comparisons can also be enabled by our preceding theoretical understanding. Specifically, our objective is to

construct confidence intervals $\{\mathcal{I}_i\}_{i\in[n]} \subseteq [n]$ for the estimated ranking of a single column of the left factor $\mathbf{X}^*$ such that

$$\mathbb{P}[\text{ranking}(X^*_{i,1}) \in \mathcal{I}_i = \{l_i, l_i+1, \cdots, u_i-1, u_i\},\ \forall i \in [n]] \le 1-\alpha.$$

A natural approach to construct such confidence intervals is to examine the distribution of the pairwise difference statistic, as also studied in Fan et al. (2025c) for the BTL models:

$$\mathcal{T} := \max_{i\neq j\in[n]} \frac{\big|\widehat{X}_{i,1} - \widehat{X}_{j,1} - X^*_{i,1} + X^*_{j,1}\big|}{\big(\sigma^2_{X_{i,1}} + \sigma^2_{X_{i,2}}\big)^{\frac{1}{2}}}. \tag{6}$$

As it fits our Gaussian multiplier bootstrap framework, the corresponding resampling form is given by

$$\max_{i\neq j\in[n]} \frac{\big|\Delta^{[b]}_{X_{i,1}} - \Delta^{[b]}_{X_{i,1}}\big|}{\big(\widehat{\sigma}^2_{X_{i,1}} + \widehat{\sigma}^2_{X_{i,1}}\big)^{\frac{1}{2}}}, \tag{7}$$

where $\widehat{\sigma}_{X_{i,1}}$ represents a suitable plug-in estimator for $\sigma_{X_{i,1}}$. As the resampling procedure returns us an estimate $\widehat{q}_\alpha$ for the α-quantile of $\mathcal{T}$, the confidence bands for the rankings is contructed according to how the oscillation ranges of the estimates overlap:

$$\mathcal{I}_i := \Big\{l \in [n] : \mathcal{J}_l \cap \mathcal{J}_i \neq \emptyset\Big\}, \quad \text{where } \mathcal{J}_l := \big[\widehat{X}^*_{l,1} - \widehat{q}_\alpha\big(\widehat{\sigma}^2_{X_{l,1}} + \widehat{\sigma}^2_{X_{l,1}}\big)^{\frac{1}{2}}, \widehat{X}^*_{l,1} + \widehat{q}_\alpha\big(\widehat{\sigma}^2_{X_{l,1}} + \widehat{\sigma}^2_{X_{l,1}}\big)^{\frac{1}{2}}\big].$$

4 Procedures Before Gradient Descent

As noted repeatedly, a fundamental prerequisite for the trajectory analysis is that the initialization lies within the RIC. In linear latent factor models, a spectral initialization on data $\mathbf{R}$ is often adequate (Ma et al., 2018), in merit of recent developments in entrywise eigenvector and singular subspace perturbation theory (Abbe et al., 2020; Agterberg et al., 2022; Cai et al., 2021). In contrast, for our generalized model, a direct application of truncated SVD on the groundtruth matrix $\psi(\mathbf{1}\boldsymbol{\zeta}^{*\top} + \mathbf{X}^*\mathbf{Y}^{*\top})$, fails to inform information of latent factors $\mathbf{X}^*$ and $\mathbf{Y}^*$, subject to the entrywise nonlinear transformation. Indeed, it is direct to see that $\psi(\mathbf{1}\boldsymbol{\zeta}^{*\top} + \mathbf{X}^*\mathbf{Y}^{*\top})$ does not even preserve the rank of low-rank structure in most cases.

Nevertheless, when the link function ψ is known, this mismatch can be partially overcome through a two-step SVD procedure (Ma et al., 2020; Zhang et al., 2020), which yields factor estimates that are consistent in Frobenius norm.

Then the challenge boils downs to how to sharpen a Frobenius-norm-faithful estimator to row-wise consistency, which is called for by the *region of incoherence and contraction* in Section 2.3. To address this issue, we incorporate a one-step refinement procedure, inspired by Chen and Li (2024), to further improve the spectral initializer. Establishing the corresponding consistency guarantees requires new technical ideas to simultaneously account for missing observations and the nonlinear link.

Leave-one-out Analysis. Before proceeding to the statistical guarantees for the estimators, we briefly reiterate a main tool underlying our gradient descent analysis. A central component is the leave-one-out argument, which has been widely used to obtain fine-grained control of iterative estimators; see, for example, Abbe et al. (2020); Chen et al. (2019a). The motivation is to control terms involving the product of a single noise vector and a gradient descent iterate, for instance,

$\mathcal{P}_\Omega\big(\mathbf{R}-\psi(\mathbf{1}\boldsymbol{\zeta}^{*\top}+\mathbf{X}^*\mathbf{Y}^{*\top})\big)_{i,\cdot}\,\mathbf{Y}^t$. To this end, we introduce a surrogate $\mathbf{Y}^{t,(i)}$ for $\mathbf{Y}^t$ that is independent of $\mathcal{P}_{\boldsymbol{\Omega}}\big(\mathbf{R}-\psi(\mathbf{1}\boldsymbol{\zeta}^{*\top}+\mathbf{X}^*\mathbf{Y}^{*\top})\big)_{i,\cdot}$, thereby enabling sharp concentration via independence. The key is to ensure that $\mathbf{Y}^{t,(i)}$ is both strictly independent of the ith row observations and sufficiently close to the original iterate $\mathbf{Y}^t$. This, in turn, requires coupled constructions not only for the gradient descent iterates themselves, but also for the preceding initialization and refinement steps, which poses new technical challenges. For the spectral method, although leave-one-out analyses for random matrices with low-rank signals are well understood, extending these techniques to settings where the expectation $\mathbf{R}^*$ of the observations may be of high rank is technically challenging. Moreover, to the best of our knowledge, a systematic leave-one-out analysis for the one-step refinement procedure has not yet been developed.

4.1 First Step: Soft Singular Value Thresholding

A crucial step in launching the optimization pipeline is to construct a sensible initialization that is sufficiently close to the true parameters under an appropriate metric. Interestingly, Ma et al. (2020), followed by Zhang et al. (2020), proposed a procedure that applies SVD twice together with knowledge of the link function. Our approach is similar in spirit, but replaces the *universal singular value thresholding* (Chatterjee, 2015) with the soft singular value thresholding method developed by Koltchinskii et al. (2011), which is more amenable to the theoretical stability required in our leave-one-out analysis. The full procedures are presented in Algorithm 3.

Note that, albeit these estimates do not necessarily achieve the optimal rates for parameter estimation, they are theoretically tractable when examined under a fine-grained analysis. This tractability pertains not only to the consistency of the resulting estimators themselves, but also to their closeness to their leave-one-out variants.

---

3 Double SVD with link inversion

Input: $\mathbf{R}\in\mathbb{R}^{n\times p}$, threshold τ, truncation bounds $(\underline{C}_\psi,\bar{C}_\psi)$, sampling rate π

Output: $\widehat{\boldsymbol{\zeta}}^{\mathsf{spec}}$, $\widehat{\mathbf{X}}^{\mathsf{spec}}$, $\widehat{\mathbf{Y}}^{\mathsf{spec}}$

Compute the SVD $\mathbf{R}=\bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}^\top$, where $\bar{\boldsymbol{\Sigma}}=\mathsf{diag}(\bar{\sigma}_i)$;

Set $\lambda=C_\lambda\sigma\sqrt{\pi d}\xi$ and

$$\widehat{\mathbf{R}}^{\mathsf{spec}}=\pi^{-1}\bar{\mathbf{U}}\mathsf{diag}\big((\bar{\sigma}_i-\lambda)_+\big)\bar{\mathbf{V}}^\top.$$

Clip and invert entrywise:

$$\check{\mathbf{M}}=\psi^{-1}\Big(\mathcal{T}_{\mathcal{D}_\psi}\big(\widehat{\mathbf{R}}^{\mathsf{spec}}\big)\Big).$$

Set

$$\widehat{\boldsymbol{\zeta}}^{\mathsf{spec}}=\mathbf{1}_n^\top\check{\mathbf{M}}/n.$$

Compute the rank-r SVD

$$\check{\mathbf{M}}-\mathbf{1}_n(\widehat{\boldsymbol{\zeta}}^{\mathsf{spec}})^\top=\mathbf{L}\mathbf{D}\mathbf{R}^\top,$$

and output

$$\widehat{\mathbf{X}}^{\mathsf{spec}}=\mathbf{L}\mathbf{D}^{1/2},\qquad\widehat{\mathbf{Y}}^{\mathsf{spec}}=\mathbf{R}\mathbf{D}^{1/2}.$$

While the spectrum of $\mathbb{E}[\mathbf{R}]$ is not fully understood, the advantage of this approach lies in its autonomy from specific structural assumptions on $\mathbb{E}[\mathbf{R}]$ since the thresholding level only relies on fluctuation of a noise matrix. This is particularly useful in the context of nonlinear-transformed matrices, where ranks of expectation matrices are usually inflated due to the nonlinear transforma-

tion.

Formally, we have the following theorem, whose proof is deferred to the Supplementary.

Theorem 4.1 (Spectral Initialization)**.** *Instate the assumptions in 2.1. Then, with probability at least* $1 - d^{-c}$*, we have*

$$\big\|\widehat{\boldsymbol{\zeta}}^{\mathsf{spec}} - \boldsymbol{\zeta}^*\big\|_2 \lesssim \underline{c}_\psi^{-1}\xi\sqrt{d/\pi}/\sqrt{n} =: \xi_{\boldsymbol{\zeta}}^{\mathsf{spec}},$$
$$\rho^{-\frac{1}{4}}\mathsf{dist}_F(\widehat{\mathbf{X}}^{\mathsf{spec}}, \mathbf{X}^*) \vee \rho^{\frac{1}{4}}\mathsf{dist}_F(\widehat{\mathbf{Y}}^{\mathsf{spec}}, \mathbf{Y}^*) \lesssim \kappa^3\underline{c}_\psi^{-1}\xi\sqrt{d/\pi}\rho^{\frac{1}{4}}/(\sigma_r^*)^{\frac{1}{2}} =: \xi^{\mathsf{spec}}.$$

Under mild conditions, one can deduce that the spectral method provides us with consistent estimates for $\boldsymbol{\zeta}^*$, $\mathbf{X}^*$, and $\mathbf{Y}^*$. Compared with regularized likelihood-based approaches (Davenport et al., 2014; Fan et al., 2025b), the spectral method sacrifices statistical efficiency due to its disregard of the underlying parametric structure, resulting in larger estimation errors.

However, the existing theoretical guarantees for Algorithm 3 — namely, consistency in the Frobenius norm—are insufficient for our purposes. In particular, they do not provide the row-wise error control required for a rigorous gradient descent analysis under partial observations. To address this challenge, we enhance the spectral estimator through an alternating-minimization refinement step prior to initializing the gradient descent procedure.

4.2 Second Step: One-Step Refinement

The *one-step refinement* (OS) algorithm is a widely used optimization technique for solving non-convex problems. For the generalized latent factor model, the magic of the OS algorithm lies in its improvement from a Frobenius-norm-consistent estimate to a row-wise consistent estimate. The OS algorithm is stated as follows: The minimization of the above optimization problems can be

---

4 Unileteral Refinement (OS)

Input: Initial estimates $\widehat{\boldsymbol{\zeta}}$, $\widehat{\mathbf{X}}$, $\widehat{\mathbf{Y}}$
Output: $\widehat{\boldsymbol{\zeta}}^{\mathsf{os}}$, $\widehat{\mathbf{X}}^{\mathsf{os}}$, $\widehat{\mathbf{Y}}^{\mathsf{os}}$
Initialize by trimming the input

$$\boldsymbol{\zeta}^0 = \mathcal{P}_\iota(\widehat{\boldsymbol{\zeta}}), \qquad \mathbf{X}^0 = \mathcal{P}_\iota(\widehat{\mathbf{X}}), \qquad \mathbf{Y}^0 = \mathcal{P}_\iota(\widehat{\mathbf{Y}}).$$

Update $\mathbf{X}$ by

$$\widehat{\mathbf{X}}^{\mathsf{os}} = \arg\min_{\mathbf{X}} \sum_{(i,j)\in\Omega} \Big[-R_{ij}\big(\zeta_j^0 + \mathbf{x}_i^\top\mathbf{y}_j^0\big) + \psi\big(\zeta_j^0 + \mathbf{x}_i^\top\mathbf{y}_j^0\big)\Big].$$

Update $(\boldsymbol{\zeta}, \mathbf{Y})$ by

$$(\widehat{\boldsymbol{\zeta}}^{\mathsf{os}}, \widehat{\mathbf{Y}}^{\mathsf{os}}) = \arg\min_{\boldsymbol{\zeta},\mathbf{Y}} \sum_{(i,j)\in\Omega} \Big[-R_{ij}\big(\zeta_j + \widehat{\mathbf{x}}_i^{\mathsf{os}\top}\mathbf{y}_j\big) + \psi\big(\zeta_j + \widehat{\mathbf{x}}_i^{\mathsf{os}\top}\mathbf{y}_j\big)\Big].$$

---

solved efficiently using first/second-order optimization methods since the problems themselves are convex with respect to the parameters $\boldsymbol{\zeta}$, $\mathbf{X}$, and $\mathbf{Y}$. The intuition behind the rowwise refinement is that, focusing on the i-th row of the latent factor, the optimizer $\widehat{\mathbf{x}}_i^{\mathsf{os}}$ is able to decouple the impact of the i-th row noise from the other rows, which enables tighter control on the rowwise consistency. This improvement is also validated by numerical experiments in Section 6.

We note that, to enable row-wise improvement over initializations in the presense of intercept, we are supposed to update $\mathbf{Y}$ and $\boldsymbol{\zeta}$ simultaneously to overcome technical issues. Nonetheless, solving the optimization problem with respect to two variables does not pose a significant challenge, as the problem is still jointly convex, given the approximate orthogonality between $\widehat{\mathbf{X}}^{\mathsf{os}}$ and $\mathbf{1}_n$.

Formally, we have the theorem below whose proof is presented in the Supplementary.

Theorem 4.2. *For Algorithm 4 combined with preceding estimates from Algorithm 3, we instate the assumptions in Theorem 2.1. Then with probability at least $1 - O(d^{-c})$ with some constant c, one has*

$$\big\|\widehat{\boldsymbol{\zeta}}^{\mathsf{os}} - \boldsymbol{\zeta}^*\big\|_2 \lesssim \kappa\kappa_\psi \xi_{\boldsymbol{\zeta}}^{\mathsf{spec}}, \qquad \rho^{\frac{1}{4}}\mathsf{dist}_F(\widehat{\mathbf{X}}^{\mathsf{os}}, \mathbf{X}^*) \vee \rho^{-\frac{1}{4}}\mathsf{dist}_F(\widehat{\mathbf{Y}}^{\mathsf{os}}, \mathbf{Y}^*) \lesssim \kappa\kappa_\psi \xi^{\mathsf{spec}}$$

$$\big\|\widehat{\boldsymbol{\zeta}}^{\mathsf{os}} - \boldsymbol{\zeta}^*\big\|_\infty \lesssim \sigma\sqrt{\frac{\iota\kappa r \log d}{\underline{c}_\psi^2 \pi n}} + \kappa\kappa_\psi \xi \sqrt{\frac{\mu r}{p}} \xi_{\boldsymbol{\zeta}}^{\mathsf{spec}},$$

$$\rho^{-\frac{1}{4}}\mathsf{dist}_{2,\infty}(\widehat{\mathbf{X}}^{\mathsf{os}}, \mathbf{X}^*) \lesssim \sigma\sqrt{\frac{\iota\kappa r \log d}{\underline{c}_\psi^2 \pi \sigma_r^*}} + \kappa\kappa_\psi \xi \sqrt{\frac{\mu r}{n}} \xi^{\mathsf{spec}},$$

$$\rho^{\frac{1}{4}}\mathsf{dist}_{2,\infty}(\widehat{\mathbf{Y}}^{\mathsf{os}}, \mathbf{Y}^*) \lesssim \sigma\sqrt{\frac{\iota\kappa r \log d}{\underline{c}_\psi^2 \pi \sigma_r^*}} + \kappa\kappa_\psi \xi \sqrt{\frac{\mu r}{p}} \xi^{\mathsf{spec}}.$$

An immediate implication of the above characterization is a locally concentrated error bound: even when $\mathsf{dist}_{2,\infty}(\widehat{\mathbf{X}}, \mathbf{X}^*)$ is large, the error can be dispersed across all rows, at the cost of only additional ξ, κ, κ_ψ, and logarithmic factors. This, in turn, allows us to identify a rather mild condition under which the output of the first two steps enters the *region of incoherence and contraction*.

Remark 6. *Note that, the bounds provided in Chen and Li (2024) shows a difference between the output with sample spliting and the one without sample splitting by a factor of $\pi^{-\frac{1}{2}}$, which is crucial when the sampling rate is small (one may think of the information-theoretic minimum sampling rate $\pi \asymp \frac{1}{n \wedge p}\mathsf{polylog}(d)$). However, they also admitted that in their simulations, the method without sample splitting performs even better. This inconsistency is because, taking our setting for example, dependency between noise and our estimate arises when dealing with $\mathbf{E}_{i,\cdot}\mathsf{diag}(\Omega_{i,\cdot})\widehat{\mathbf{Y}}^{\mathsf{spec}}$, which need to be decoupled using sample spliting in their proof. A similar difficulty also appears in the proof of Chernozhukov et al. (2023b). Nevertheless, as noted in their discussion, this dependence can instead be handled through a leave-one-out argument. This observation coincides our analysis, which avoids the additional $\pi^{-1/2}$ loss induced by sample splitting.*

5 Related Literature

The literature on latent factor models is vast, and we refer the readers to Bartholomew et al. (2008) for a comprehensive overview. In terms of exploiting low-rank structures to faclitate the estimation, we refer readers to Candes and Recht (2012); Chen et al. (2019a, 2020, 2019b); Chen and Li (2024) for a series of works on linear latent factor models. One-bit matrix completion problem: Davenport et al. (2014); Cai and Zhou (2013)

Inferential results on latet low-rank structure have been well studied in the linear settings, to name a few, Chen et al. (2019b); Xie and Xu (2023); Chernozhukov et al. (2023b). The understanding of nonlinear structure is still in its infancy. Further, Li et al. (2023) studies the asymptotic behavior of the regularized MLE for symmetric latent space models, which similarly involves link functions but requires additional regularizations, where a subsequent work (Ouyang et al., 2024) extends the results to generalized latent factor models with covariates. However, the aforementioned

works merely focus on the statistical aspects of models, but fall short of providing a comprehensive computational framework with theoretical guarantees.

The heuristic of our approach relies on the implicit regularization of gradient descent, which has been extensively studied in the context of linear latent factor models (Chen et al., 2020, 2019b; Ma et al., 2018). The recent work (Fan et al., 2025a) extends the results to generalized latent factor models with covariates, but it does not address the missingness issue. Our work fills this gap by providing a comprehensive framework for inference in generalized latent factor models with missing data.

6 Numerical Simulations

This section studies the finite-sample behavior of the proposed generalized latent factor pipeline. We first examine the stability of the final estimator under varying signal strength and missingness, in comparison with the spectral initializer, the one-step refinement. Then, we examine the inferential procedures through comparing the estimated distributions and the ones from Monte-Carlo sampling.

6.1 Simulation Setup

Consider the logistic latent factor model

$$R_{ij} \mid \Theta^*_{ij} \sim \text{Bernoulli}(\sigma(\Theta^*_{ij})), \qquad \Theta^*_{ij} = \zeta^*_j + x_i^{*\top} y^*_j,$$

where each entry is observed independently with probability π. The left latent facor is generated from $\rho\sqrt{np} \cdot \bar{\mathbf{U}}\mathbf{E}_U \cdot \mathsf{diag}(\sqrt{3/2}, 1)$, where $\mathbf{V} \in \mathcal{O}(n, n-1)$ is an arbitrary orthonormal matrix with $\mathbf{1}_n^\top \bar{\mathbf{U}} = \mathbf{0}$ and $\mathbf{E}_U$ is sampled from $\mathsf{Unif}(\text{Stiefel}_{r,n-1})$, while the right latent factor follows $\rho\sqrt{np} \cdot \mathbf{E}_V \cdot \mathsf{diag}(\sqrt{3/2}, 1)$ with $\mathbf{E}V \sim \mathsf{Unif}(\text{Stiefel}_{r,p})$. The resulting matrix $\mathbf{X}^*\mathbf{Y}^{*\top}$ is then rescaled to have a Frobenius norm of $\rho\sqrt{np}$, and then rescaled through the top-r singular value decomposition so that the identifiability constraints are satisfied. We set $\zeta^* = -0.6\,\mathbf{1}_p$, use $\pi = 0.5$ as the baseline sampling rate, and take $\eta = 0.003$ as the baseline gradient step size.

6.2 Stability Under Signal, Missingness, and Learning Rate

We begin with the full estimator. The goal of this experiment is to identify the practical stability region of the pipeline after all three stages have been applied. We set $n = 1000$, $p = 500$, and $r = 2$, and vary the signal strength $\lambda \in \{0.3, 0.4, \cdots, 1.0\}$ and the sampling rate $\pi \in \{0.1, 0.2, 0.35, 0.5, 0.7, 0.9\}$. Figure 1 plots the mean relative errors against the signal strengthλ and the sampling rate π, with separate panels for π and separate curves for λ.

Across 200 replications, the relative Frobenius errors decrease as either the signal strength λ or the observation probability π increases. The spectral initializer provides a reasonable starting point, while the one-step refinement yields a substantial improvement and the final GD step gives an additional, more modest reduction. The improvement is especially pronounced as the observation rate increases, indicating that denser observations substantially stabilize both model-side and task-side factor recovery. Overall, the results confirm that the proposed multi-stage procedure consistently improves latent factor estimation over the spectral baseline and behaves as expected under stronger signal and more complete observation.

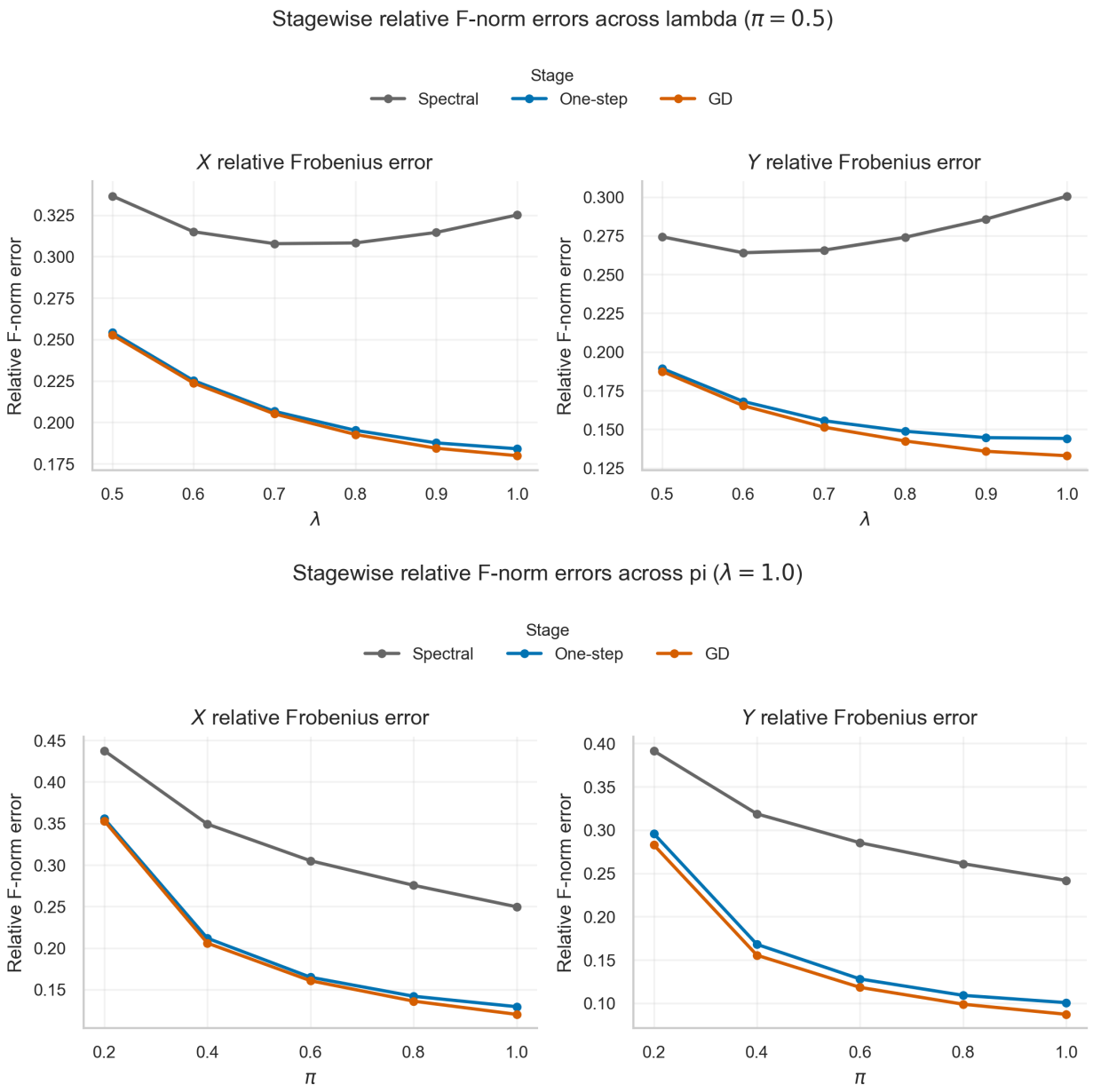

Figure 1: Estimation errors across algorithmic stages under varying signal strength λ and sampling rate π.

6.3 Asymptotics of Estimates

Finally, we set $n = p = 2000$ and $\pi = 0.5$ to examine the asymptotic normality of the entries as well as the consistency of the bootstrapping procedures. We start with the asymptotic normality regarding $(\mathbf{X}^*)_{1,1}$, whose behaviour is suggested in Theorem 3.2, then turn to checking the distributions of the ranking statistic (6) as well as the bootstrap counterpart (7).

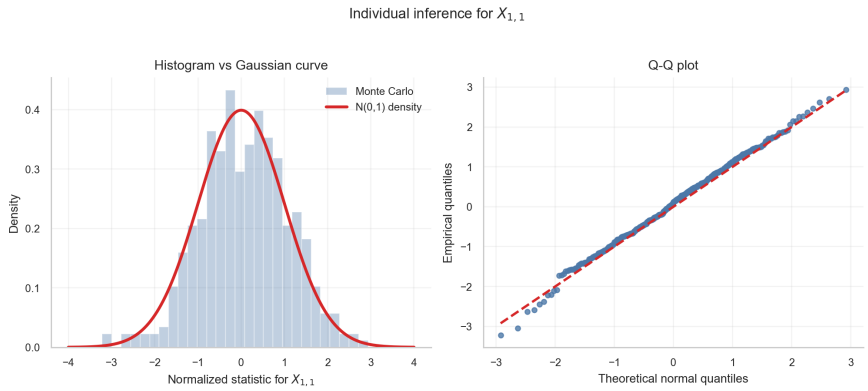

Figure 2: Histogram and Q-Q Diagnostic for Individual Normal Approximation

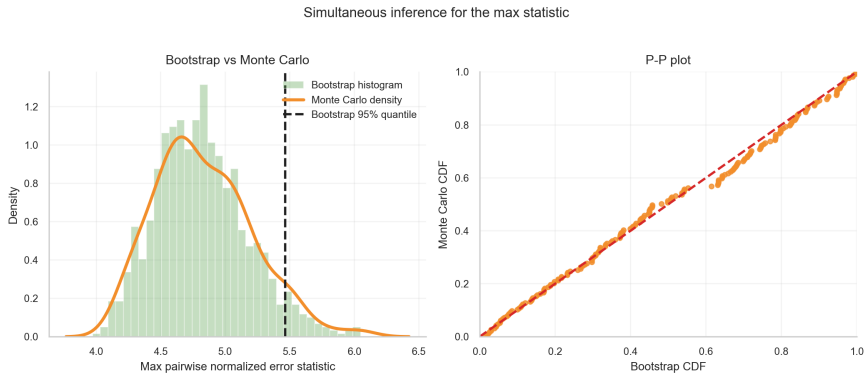

Figure 3: Bootstrap and Monte Carlo Calibration of the Simultaneous Ranking Statistic

From Figure 2, we can see that the limiting distribution of (6) is approximately standard normal, which indicates that the gradient descent algorithm is able to achieve the asymptotic normality. And through Figure 3, we observe that the Gaussian multiplier bootstrap is able to mimic the limiting distribution of the targeted statistic, which indicates that the Gaussian multiplier bootstrap is able to provide valid confidence bands for the estimates.

7 Conclusion and Discussion

We have introduced a novel and provable three-step algorithm for uncertainty quantification in generalized latent factor models with missing binary data. First, we obtain a Frobenius-consistent initialization via a double-SVD procedure. Next, we refine these estimates through an alternating-minimization procedure to achieve rowwise consistency. Finally, we apply vanilla gradient descent—benefiting from its implicit regularization–to locate an approximate stationary point of the non-convex likelihood.

Our heuristic and empirical results confirm that this workflow delivers both accurate point estimates and valid confidence bands. The linear approximation of the gradient-descent iterates underpins our inferential framework, allowing us to characterize their limiting distributions and to construct individual and simultaneous confidence sets via the Gaussian multiplier bootstrap. Moreover, by analyzing the loss landscape within the Region of Incoherence and Contraction, we show why gradient descent remains robust to local minima and permits aggressive step sizes.

Our simulations demonstrate that the proposed method achieves high estimation precision, reliable coverage, and resilience to outliers. This work thus provides a practical and theoretically grounded toolkit for inference in high-dimensional latent factor settings with missingness.

Several paths are left open for further work: (i) such nonlinearity-encoded factor structure is also natural for symmetric network data as well as tensor data, thus it would be interesting to see whether we can have computationally tractable and statistically optimal procedures for such models; (ii) While our analysis allows the minimum sampling rate in the balanced regime ($n \approx p$), the singular subspace perturbation theory suggests that, in the unbalanced regime ($n \gg p$ or $n \ll p$), the error rates for the left and right factors should scale according to their respective dimensions. Characterizing the sharp dependence on these dimensions in the present nonlinear setting is an important question for future study.

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452.

Agterberg, J., Lubberts, Z., and Priebe, C. E. (2022). Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *IEEE Transactions on Information Theory*, 68(7):4618–4650.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension.

Bartholomew, D. J., Steele, F., and Moustaki, I. (2008). *Analysis of multivariate social science data*. CRC press.

Cai, C., Li, G., Chi, Y., Poor, H. V., and Chen, Y. (2021). Subspace estimation from unbalanced and incomplete data matrices: statistical guarantees. *The Annals of Statistics*, 49(2):944–967.

Cai, T. and Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res.*, 14(1):3619–3647.

Candes, E. and Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding.

Chen, J., Liu, D., and Li, X. (2020). Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 66(9):5806–5841.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2019a). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *arXiv preprint arXiv:1902.07698*.

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019b). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937.

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2021). Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *Annals of statistics*, 49(5):2948.

Chen, Y. and Li, X. (2024). A note on entrywise consistency for mixed-data matrix completion. *Journal of Machine Learning Research*, 25(343):1–66.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors.

Chernozhukov, V., Chetverikov, D., and Koike, Y. (2023a). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *The Annals of Applied Probability*, 33(3):2374–2425.

Chernozhukov, V., Hansen, C., Liao, Y., and Zhu, Y. (2023b). Inference for low-rank models. *The Annals of statistics*, 51(3):1309–1330.

Chernozhuokov, V., Chetverikov, D., Kato, K., and Koike, Y. (2022). Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5):2562–2586.

Davenport, M. A., Plan, Y., Van Den Berg, E., and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223.

Fan, J., Fan, Y., Lv, J., Yang, F., and Yu, D. (2025a). Asymptotic theory of eigenvectors for latent embeddings with generalized laplacian matrices. *arXiv preprint arXiv:2503.00640*.

Fan, J., Ge, J., and Hou, J. (2025b). Covariates-adjusted mixed-membership estimation: A novel network model with optimal guarantees. *arXiv preprint arXiv:2502.06671*.

Fan, J., Lou, Z., Wang, W., and Yu, M. (2025c). Ranking inferences based on the top choice of multiway comparisons. *Journal of the American Statistical Association*, 120(549):237–250.

Fan, J., Wang, K., Zhong, Y., and Zhu, Z. (2021). Robust high dimensional factor models with applications to statistical machine learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(2):303.

Huber, P., Ronchetti, E., and Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(4):893–908.

Kidzinski, L., Hui, F. K., Warton, D. I., and Hastie, T. J. (2022). Generalized matrix factorization: efficient algorithms for fitting generalized linear latent variable models to large data arrays. *Journal of Machine Learning Research*, 23(291):1–29.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion.

Li, J., Xu, G., and Zhu, J. (2023). Statistical inference on latent space models for network data. *arXiv preprint arXiv:2312.06605*.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2018). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR.

Ma, Z., Ma, Z., and Yuan, H. (2020). Universal latent space model fitting for large networks with edge covariates. *Journal of Machine Learning Research*, 21(4):1–67.

Ouyang, J., Cui, C., Tan, K. M., and Xu, G. (2024). Statistical inference for covariate-adjusted and interpretable generalized factor model with application to testing fairness. *arXiv preprint arXiv:2404.16745*.

Park, D., Kyrillidis, A., Caramanis, C., and Sanghavi, S. (2018). Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences*, 11(4):2165–2204.

Rohe, K. and Zeng, M. (2023). Vintage factor analysis with Varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1037–1060.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Soltanolkotabi, M., Stöger, D., and Xie, C. (2023). Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5140–5142. PMLR.

Su, L. and Wang, F. (2025). Inference for large dimensional factor models under general missing data patterns. *Journal of Econometrics*, 250:106022.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. (2016). Low-rank solutions of linear matrix equations via procrustes flow. In *International conference on machine learning*, pages 964–973. PMLR.

Wang, F. (2022). Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions. *Journal of econometrics*, 229(1):180–200.

Xie, F. and Xu, Y. (2023). Efficient estimation for random dot product graphs via a one-step procedure. *Journal of the American Statistical Association*, 118(541):651–664.

Zhang, H., Chen, Y., and Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, 85(2):358–372.

Zheng, Q. and Lafferty, J. (2016). Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*.