CRYSTAL: **C**luster-geomet**ry**-inspired **S**pec**t**ral **Al**gorithm for Binary Latent Feature Models

Yuqi Gu Chengzhu Huang Zhiyu Xu

Department of Statistics, Columbia University

Abstract

Binary latent feature models arise in many statistical settings, including educational cognitive diagnosis, recommendation systems, ecology, and overlapping community detection. Each observational unit may simultaneously possess multiple latent features, leading to up to 2^K distinct latent profiles given K binary features. This combinatorial structure creates prohibitive computational and statistical challenges, and existing likelihood-based or Bayesian methods struggle with even moderate latent dimensions. We propose CRYSTAL (Cluster-geometry-inspired Spectral Algorithm), a scalable and provable spectral method that exploits an interesting geometry: the population singular subspace embeddings of the exponentially many latent profiles form a highly structured "crystal" shape of a parallelotope, a higher-dimensional analogue of a parallelogram. Exploiting this geometry, we reduce the recovery of latent profiles to identifying a small number of geometrically distinguished directions in a low-dimensional space. Theoretically, we establish identifiability for models with and without intercepts. In high-dimensional regimes, we prove that CRYSTAL exactly recovers the latent feature matrix with high probability, and we derive nonasymptotic error bounds for parameter estimation. We further prove asymptotic normality of the spectral estimators and develop plug-in covariance estimators for statistical inference. Extensive simulations corroborate the theory, and applications to educational assessment and ecology data illustrate the method's interpretability and scalability.

Keywords: Binary Latent Feature Models; Overlapping Clustering; Spectral Methods; Exact Recovery; High-dimensional Inference; Cognitive Diagnosis Models.

1 Introduction

Binary Latent Feature Models and Overlapping Clustering. Binary latent features arise in a wide range of statistical problems in which each observational unit may simultaneously possess multiple latent attributes or belong to multiple latent groups. Examples include cognitive diagnosis models in educational assessment (Chen et al., 2015; Xu

The authors' names are ordered alphabetically. Emails: {yuqi.gu, ch3786, zx2488}@columbia.edu.

and Shang, 2018), where each student test-taker has a mastery/deficiency profile of multiple latent skills; recommendation systems with overlapping user preferences (Heller and Ghahramani, 2007); ecological studies in which species exhibit multiple habitat affinities (Scherting and Dunson, 2024; Zhou et al., 2025); and network and text analysis with overlapping clustering structures (Latouche et al., 2011; Zhang et al., 2020). Statistically, these problems can be naturally described by *binary latent feature models*, in which each observational unit carries a binary latent vector $\mathbf{a}_i = (a_{i1}, \ldots, a_{iK}) \in \{0,1\}^K$ and the observed measurements depend on the active latent features. In these settings, the number of distinct latent profiles grows exponentially with the latent dimension K.

Let $\mathbf{R} \in \mathbb{R}^{N \times J}$ denote a data matrix with rows indexing observational units $i \in [N]$ and columns indexing measurements $j \in [J]$. We consider binary latent feature models in which $\mathbb{E}(R_{ij}) \equiv R^*_{ij} = d_j + \mathbf{a}_i^\top \boldsymbol{\theta}_j$, where $\mathbf{A} = (\mathbf{a}_1^{*\top}, \ldots, \mathbf{a}_N^{*\top})^\top \in \{0,1\}^{N \times K}$ is an unknown binary membership matrix, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^{*\top}, \ldots, \boldsymbol{\theta}_J^{*\top})^\top \in \mathbb{R}^{J \times K}$ is an unknown loading parameter matrix, and $\mathbf{d}^* \in \mathbb{R}^J$ is an intercept vector. The intercept-free case corresponds to $\mathbf{d}^* \equiv \mathbf{0}$, which characterizes an interesting special case of *overlapping clustering*. We allow R_{ij} to follow general observation models with expectation specified above, including Bernoulli and Poisson responses, and more generally independent noise models with mild tail control.

Challenges of Existing Methods and Gaps in the Literature. Despite its seemingly simple appearance, learning the binary latent features from data presents fundamental statistical and computational challenges. The rows of $\mathbf{A}$ contain up to $|\{0,1\}^K| = 2^K$ distinct attribute patterns (i.e., 2^K mixture components if viewed from a mixture model perspective), rendering likelihood-based estimation combinatorial in nature. Classical marginal maximum likelihood via EM algorithms and Bayesian inference methods (Heller and Ghahramani, 2007; Doshi-Velez and Ghahramani, 2009; Frank et al., 2012; Chen et al., 2015; Xu and Shang, 2018; Ni et al., 2020; Gu and Dunson, 2023) become computationally prohibitive or unstable even for moderate K. Moreover, theoretical guarantees for these methods are typ-

ically limited to consistency or identifiability in low-dimensional regimes (e.g., Gu and Xu, 2020; Chen et al., 2020a), with little understanding of finite-sample behavior or uncertainty quantification in high dimensions.

It is worth highlighting one particular application area of binary latent feature models that motivates this study: educational cognitive diagnosis modeling (CDM; Rupp and Templin, 2008; Chen et al., 2017; von Davier and Lee, 2019). In this application, each observation is a student's multivariate correct or incorrect responses to many test questions in an educational assessment, and the binary latent features are each student's latent skill profile $\mathbf{a}_i \in \{0,1\}^K$ that characterizes their mastery or deficiency statuses on K latent skills. In this context, accurately estimating students' skill profiles is of great practical importance, as this can provide fine-grained diagnoses of students' strengths and weaknesses to facilitate personalized instructions. However, most existing methods for estimating CDMs suffer from the issue of exponential and combinatorial complexity with respect to K (e.g., Chen et al., 2015, 2017; Xu and Shang, 2018; Gu and Xu, 2019; Chen et al., 2020a; Zhang et al., 2023). Furthermore, existing methods do not come with theoretical guarantees for accurate skill profile estimation or statistical inference on parameters in high dimensions, which are highly desirable features to enable trustworthy educational measurement and intervention.

On the other hand, spectral methods (e.g., Chen et al., 2021) have emerged as powerful tools for estimating latent variable models. However, existing spectral methods are primarily developed for non-overlapping clustering (Löffler et al., 2021; Zhang and Zhou, 2024; Lyu et al., 2025) or mixed-membership models with simplex constraints (Zhang et al., 2020; Jin et al., 2024; Mao et al., 2021; Ke and Wang, 2024). Their extension to binary latent feature models with intercepts remains theoretically underexplored.

New Geometric Insight of a Parallelotope in the Spectral Domain. We uncover a new geometric insight for binary latent feature models that transforms the combinatorial recovery problem into a structured *spectral estimation* task. At the population level, we

show that the leading K *singular subspace* of the data matrix $\mathbf{R}$ exhibits a striking geometry: the spectral embeddings of the exponentially many binary vector configurations lie on the vertices of a highly structured *parallelotope*, which is a higher-dimensional analogue of a parallelogram. In particular, the K principal edges of this parallelotope correspond to pure latent patterns (i.e., data points with latent binary vectors $\mathbf{a}_i$ equal to the canonical basis vectors), and all other multi-attribute patterns arise as sums along the principal edges; see Figure 1. This geometry is distinct from the simplex structure leveraged in mixed-membership models (e.g., Jin et al., 2024; Ke and Wang, 2024; Mao et al., 2021; Zhang et al., 2020; Chen et al., 2024): overlap produces a parallelotope rather than a simplex.

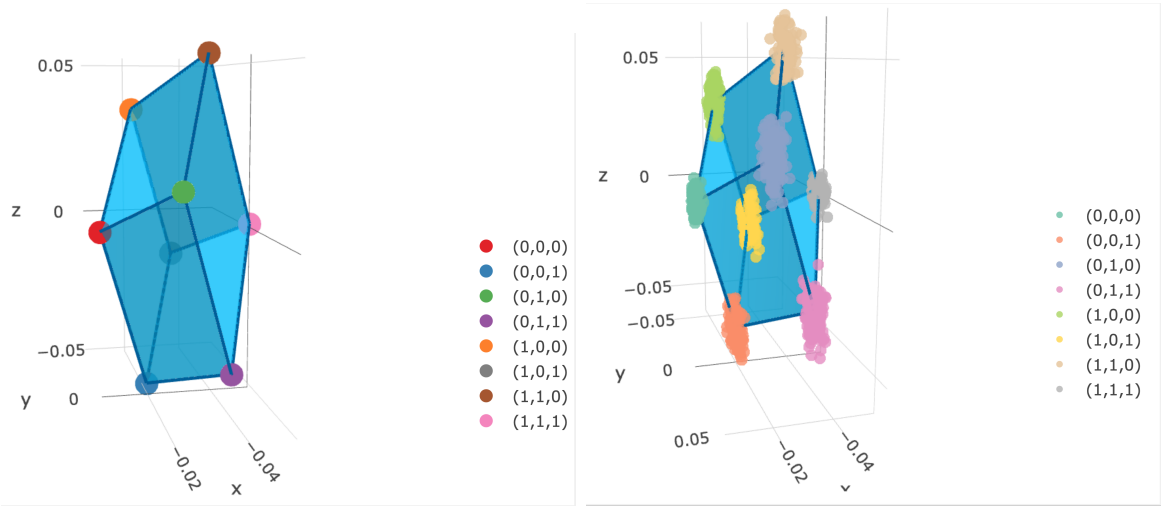

Figure 1: "Crystal" geometry of the left singular subspace embeddings, in the noiseless population case (left panel) and the noisy sample case (right panel).

The "crystal" geometry implies that recovering exponentially many latent attribute profiles does not require searching over 2^K configurations. Instead, we show that it suffices to identify the origin of the parallelotope together with K geometrically distinguished directions in the K-dimensional singular subspace. This insight motivates our estimation procedure.

Methodological Contribution: CRYSTAL. Building on the above geometry, we propose CRYSTAL (**C**luster-geometry-inspired **S**pec**t**ral **Al**gorithm), a scalable and provable spectral method for estimating binary latent feature models. We develop different versions

of the algorithm for models with and without intercepts. For models *without intercepts*, we first extract the leading singular vectors of the observed data matrix that expose the approximate parallelotope geometry under noise. Then, we identify pure latent patterns via a successive projection algorithm (Gillis and Vavasis, 2013), and recover the full binary latent feature matrix by solving a linear system followed by hard thresholding. For models *with intercepts*, we develop an augmented procedure that first identifies rows corresponding to the pure-intercept latent patterns and then prunes their effect before applying the same geometric recovery strategy to the remaining subspace. Importantly, CRYSTAL avoids likelihood maximization, combinatorial optimization, and MCMC sampling-based inference. Its computational complexity scales linearly in the latent dimension K, making it scalable for high-dimensional applications where existing methods are infeasible.

Theoretical Contributions. We provide a unified and comprehensive statistical theory for recovery, estimation, and inference in binary latent feature models. First, we establish identifiability of the models both with and without intercept terms, under interpretable conditions. Second, in high-dimensional regimes where both N and J go to infinity, we derive nonasymptotic exact recovery guarantees for individuals' latent attribute profiles $\mathbf{A}^*$, showing that CRYSTAL recovers the true $\mathbf{A}^*$ with high probability under explicit signal-to-noise-ratio and spectral gap conditions. Third, we derive convergence rates for estimating loading parameters $\mathbf{\Theta}^*$ (and $\mathbf{d}^*$ when intercepts are present) under general noise models, including Bernoulli and Poisson data. Finally, we prove asymptotic normality of the spectral estimators for $\mathbf{\Theta}^*$ and provide plug-in covariance estimators that enable entrywise and joint hypothesis tests; e.g., testing whether a specific loading is zero. Applied to cognitive diagnostic models in particular, these entrywise and joint loading-support tests provide a principled, data-driven route to estimating the Q-matrix (the key binary matrix describing which items load on which latent skills, Tatsuoka, 1983) under the exact-recovery regime, addressing a practical challenge in the CDM literature without requiring a pre-specified oracle Q-matrix

or restrictive sparsity assumptions on the Q-matrix. In summary, the above results support scalable estimation and principled uncertainty quantification in settings where existing approaches are computationally intractable and do not provide comparable guarantees in the high-dimensional regime.

Numerical studies and applications. Extensive simulation studies demonstrate that CRYSTAL scales favorably with K and achieves accurate recovery across Bernoulli and Poisson settings, including regimes beyond the practical computational limits of EM-based MMLE implementations. We further demonstrate the method on an international educational assessment dataset for cognitive diagnosis and an ecological abundance dataset for overlapping clustering, where the recovered latent structures are interpretable andd align with substantive domain knowledge.

Organization. Section 2 establishes identifiability and reveals the parallelotope geometry. Section 3 proposes the CRYSTAL methods for models with and without intercepts. Section 4 proves exact recovery guarantees and finite-sample error bounds for parameter estimation. Section 5 develops distributional theory and inference procedures. Section 6 conducts extensive simulation studies and Section 7 presents two real data applications. Finally, Section 8 gives concluding remarks. The proofs are included in the Supplementary Material.

2 Identifiability and Parallelotope Geometry

We first formalize identifiability for the population expectation matrix and clarify how it connects to the noisy recovery guarantees presented later. We observe a data matrix $\mathbf{R} \in \mathbb{R}^{N\times J}$ with expectation

$$\mathbb{E}[\mathbf{R}] = \mathbf{R}^* = \mathbf{A}^* \cdot \mathbf{\Theta}^{*\top} + \mathbf{1}_{N\times 1}\mathbf{d}^{*\top}. \tag{1}$$

where $\mathbf{A}^* \in \{0,1\}^{N\times K}$ encodes N individuals' K binary latent features, and $\mathbf{\Theta}^* \in \mathbb{R}^{J\times K}$ is a loading parameter matrix. This model is invariant to a simultaneous permutation of latent features: for any permutation matrix $\mathbf{\Pi} \in \mathbb{R}^{K\times K}$, $\mathbf{A\Theta}^\top = (\mathbf{A\Pi})(\mathbf{\Theta\Pi})^\top$. Accordingly,

identifiability can at best hold up to a permutation of the K latent dimensions.

Definition 1 (Identifiability up to permutation)**.** *Consider a parameter class $\mathcal{P}$ of triples $(\mathbf{A}, \boldsymbol{\Theta}, \mathbf{d})$ satisfying structural constraints (e.g., $\mathbf{A} \in \{0,1\}^{N\times K}$ and anchor/pure rows exist). We say the model (1) is identifiable over $\mathcal{P}$ up to permutation if whenever*

$$\mathbf{A}\boldsymbol{\Theta}^\top + \mathbf{1}\mathbf{d}^\top = \bar{\mathbf{A}}\bar{\boldsymbol{\Theta}}^\top + \mathbf{1}\bar{\mathbf{d}}^\top, \qquad (\mathbf{A}, \boldsymbol{\Theta}, \mathbf{d}),\ (\bar{\mathbf{A}}, \bar{\boldsymbol{\Theta}}, \bar{\mathbf{d}}) \in \mathcal{P},$$

then $\bar{\mathbf{d}} = \mathbf{d}$ and there exists a permutation matrix $\boldsymbol{\Pi}$ such that $\bar{\mathbf{A}} = \mathbf{A}\boldsymbol{\Pi}$ and $\bar{\boldsymbol{\Theta}} = \boldsymbol{\Theta}\boldsymbol{\Pi}$.

We emphasize that the above identifiability notion differs from both the population identifiability considered in Chen et al. (2015), Xu and Shang (2018) and Gu and Dunson (2023) for low-dimensional latent trait models, and the structural identifiability considered in Chen et al. (2020b) and Gu and Xu (2023) in the asymptotic sense. Our identifiability notion focuses on the "expectation identifiability" (Chen et al., 2024), which is a suitable notion and useful prerequisite for consistent spectral estimation in the double-asymptotic regime with both N and J going to infinity. This identifiability notion is similar to those considered in mixed membership network models or topic models (Jin et al., 2024; Ke and Wang, 2024; Mao et al., 2021).

2.1 Identifiability of the Model Without Intercepts

We first consider the intercept-free model with expectation $\mathbb{E}[\mathbf{R}] = \mathbf{R}^* = \mathbf{A}^*\boldsymbol{\Theta}^{*\top}$.

Proposition 1 (Noiseless Identifiability of the Intercept-free Model)**.** *Suppose $\boldsymbol{\Theta}^*$ has full column rank K, and the row vectors of $\mathbf{A}^*$ contain $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$. Then in the noiseless setting, from $\mathbf{R}^* = \mathbf{A}^*\boldsymbol{\Theta}^{*\top}$, all parameters $(\mathbf{A}^*, \boldsymbol{\Theta}^*)$ are identifiable.*

Proposition 1 is a population-level uniqueness statement: it rules out multiple alternative factorizations $\mathbf{A}^*\boldsymbol{\Theta}^{*\top}$ of the same $\mathbb{E}[\mathbf{R}]$ within the model class, except for column permutation of $\mathbf{A}^*$ and $\boldsymbol{\Theta}^*$. It does *not* by itself imply that $(\mathbf{A}^*, \boldsymbol{\Theta}^*)$ can be recovered from a noisy realization $\mathbf{R}$. Recovery guarantees given noisy data require additional signal-to-noise and singular-value separation conditions, which we impose and explain in Section 3.

2.2 Identifiability of the Model With Intercepts

For binary latent feature models with intercepts, the extra parametrization flexibility introduces additional ambiguity. The parameterization is not unique under the conditions in Proposition 1; to see this, note that for any i, j, k, the transformation

$$\bar{a}_{ik} = 1 - a_{ik}, \qquad \bar{\theta}_{jk} = -\theta_{jk}, \qquad \bar{d}_j = d_j + \theta_{jk},$$

leaves the expectation $\mathbb{E}[R_{ij}]$ unchanged and gives $d_j + a_{ik}\theta_{jk} = \bar{d}_j + \bar{a}_{ik}\bar{\theta}_{jk}$, hence $\mathbf{A}\mathbf{\Theta}^\top + \mathbf{1}\mathbf{d}^\top = \bar{\mathbf{A}}\bar{\mathbf{\Theta}}^\top + \mathbf{1}\bar{\mathbf{d}}^\top$. Therefore, without sign restrictions on $\mathbf{\Theta}$, the model admits equivalent parameterizations that are *not* mere column permutations of $\mathbf{A}$ and $\mathbf{\Theta}$.

To remove this ambiguity, we adopt the following convention, which is very natural and interpretable for Bernoulli and Poisson models: the intercept represents the baseline level and each latent feature has a nonnegative additive effect. In the application of educational cognitive diagnoses, $R_{ij} = 1$ or 0 denotes whether student i responds to question j correctly, $a_{ik} = 1$ or 0 indicates whether the student possesses the kth latent skill. In this context, θ_{jk} represents the change of the correct response probability to item j of students possessing skill k versus students lacking it. So, imposing nonnegative θ_{jk} is highly interpretable and aligns with the intuition that possessing a skill should not decrease the correct response probability. This condition is indeed widely adopted across many existing studies as a "monotonicity" constraint (e.g., Gu and Xu, 2019).

Proposition 2 (Noiseless Identifiability With Intercepts)**.** *Assume (i)* $\mathbf{\Theta}^*$ *has nonnegative entries with full column rank* K*; (ii) the rows of* $\mathbf{A}^*$ *contain* $\{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_K\}$*; and (iii)* $\mathbf{d}^* > \mathbf{0}$ *componentwise. Then* $(\mathbf{A}^*, \mathbf{d}^*, \mathbf{\Theta}^*)$ *are identifiable.*

Under the conditions of Proposition 2, the intercept is identifiable at the population level via the componentwise minimum: $d_j^* = \min_{i\in[N]} R_{ij}^* = \min_{i\in[N]} \left(d_j^* + \sum_{k=1}^K a_{ik}^*\theta_{jk}^*\right)$, since $\theta_{jk}^* \geq 0$ and there exists $i \in [N]$ such that $\mathbf{a}_i^* = \mathbf{0}$. However, this noiseless argument does not directly yield a consistent estimator of $\mathbf{d}^*$ in discrete noisy models (e.g., Bernoulli data

or Poisson data), motivating our spectral identification strategy developed in Section 3 and analyzed in Section 4.

2.3 Parallelotope Geometry of the Population Singular Subspace

We now reveal the population geometry that motivates the CRYSTAL method. Consider first the intercept-free case $\mathbf{R}^* = \mathbf{A}^*\mathbf{\Theta}^{*\top}$ and its rank-K SVD denoted by $\mathbf{R}^* = \mathbf{U}^*\mathbf{\Sigma}^*\mathbf{V}^{*\top}$. For a positive integer m, write $[m] = \{1, \ldots, m\}$. Let $S = \{S_1, \ldots, S_K\} \subseteq [N]$ be indices of a set of pure latent profiles such that $\mathbf{A}^*_{S,:} = \mathbf{I}_K$.

Focusing on the rows of $\mathbf{R}^* = \mathbf{A}^*\mathbf{\Theta}^{*\top}$ indexed by S, we obtain $\mathbf{\Theta}^{*\top} = \mathbf{U}^*_S\mathbf{\Sigma}^*\mathbf{V}^{*\top}$. This implies $\boldsymbol{\theta}^{*\top}_k = \mathbf{U}^*_{S_k}\mathbf{\Sigma}^*\mathbf{V}^{*\top}$ for each $k = 1, \ldots, K$. We can further write the left singular subspace $\mathbf{U}$ as

$$\mathbf{U}^* = \mathbf{A}^*\mathbf{\Theta}^{*\top}\mathbf{V}^*\mathbf{\Sigma}^{*-1} = \mathbf{A}^*(\mathbf{U}^*_S\mathbf{\Sigma}^*\mathbf{V}^{*\top})\mathbf{V}^*\mathbf{\Sigma}^{*-1} = \mathbf{A}^*\mathbf{U}^*_S. \tag{2}$$

Now we consider the population left-singular-subspace embedding of the ith observational unit, $\mathbf{U}^*_i$, which is the ith row of $\mathbf{U}^*$. Writing out each row of (2) and plugging in the form of $\boldsymbol{\theta}^{*\top}_k = \mathbf{U}^*_{S_k}\mathbf{\Sigma}^*\mathbf{V}^{*\top}$ gives

$$\mathbf{U}^*_i = \mathbf{a}^*_i\mathbf{U}^*_S = \sum_{k=1}^{K} a^*_{ik}\mathbf{U}^*_{S_k} = \sum_{k=1}^{K} a^*_{ik}\boldsymbol{\theta}^{*\top}_k\mathbf{V}^*\mathbf{\Sigma}^{*-1}.$$

The above expression reveals a highly structured *parallelotope* geometry of the embeddings of the up to 2^K latent profiles of $\mathbf{a}_i \in \{0,1\}^K$. Specifically, the row embeddings $\{\mathbf{U}^*_{i,:} : i \in [N]\}$ lies on the vertex set of a K-dimensional parallelotope spanned by the K "edge" vectors $\{\mathbf{U}^*_{S_k,:}\}_{k=1}^K$. There are K principal vertices of this parallelotope corresponding to the "pure profiles" with $\mathbf{a}^*_i$ equal to some canonical basis vector $\mathbf{e}_1, \ldots, \mathbf{e}_K$; see Figure 1. The distance between each of these K vertices $\mathbf{U}^*_{S_k}$ to the origin $\mathbf{0}_K$ is $\left\|\boldsymbol{\theta}^{*\top}_k\mathbf{V}^*\mathbf{\Sigma}^{*-1}\right\|$. The above "crystal" geometry reduces the combinatorial search over 2^K membership patterns to a geometric problem of identifying the K principal edges, which we operationalize in Section 3.

We remark that (2) bears resemblance to the spectral representations in mixed mem-

bership models (Ke and Wang, 2024; Jin et al., 2024; Mao et al., 2021), but differs in the specific geometry and the implication for estimation procedures. In those works, the population eigenspace embedding $\mathbf{U}_i^*$ of node i can be written as $\mathbf{U}_i^* = \sum_{k=1}^K \pi_{ik}\mathbf{U}_{S_k}^*$, where $(\pi_{i1}, \ldots, \pi_{iK})$ is the ith node's mixed membership vector living in the simplex, with continuous scores $\pi_{ik} \geq 0$ and $\sum_{k=1}^K \pi_{ik} = 1$. In contrast, in the considered binary latent feature models, the singular subspace geometry differs from the mixed membership geometry with simplex constraints. Instead of obtaining a simplex in $\mathbb{R}^K$, we reveal that the up to 2^K latent attribute patterns take the shape of a structured parallelotope in $\mathbb{R}^K$. This geometry is crucial to inspiring our estimation procedure for the binary latent feature matrix $\mathbf{A}^*$ with exact recovery guarantees.

3 CRYSTAL Method

Building on the population-level understanding of $\mathbb{E}[\mathbf{R}] = \mathbf{R}^*$, we turn to its empirical counterpart $\mathbf{R}$, i.e., the data matrix. We aim to extract information from the leading left singular subspace of $\mathbf{R}$ to recover the binary latent profiles. Different procedures will be developed for cases with and without intercepts, under the respective identifiability conditions.

3.1 Algorithm for Intercept-Free Models

For models without intercepts with $\mathbb{E}[\mathbf{R}] = \mathbf{R}^* = \mathbf{A}^*\mathbf{\Theta}^{*\top}$, we propose a two-step estimation procedure inspired by the parallelotope geometry: (i) first identifying the pure latent profiles or the K geometrically distinct directions in the singular subspace, and (ii) then recovering all rows of $\mathbf{A}$ based on the preceding pure latent profile estimates.

Figure 2 illustrates the geometric properties of the overlapping clustering structure following SCORE normalization (Jin, 2015) of the singular subspace, in the case of three latent attributes. The role of SCORE normalization in CRYSTAL is to convert the additive parallelotope geometry into a separable convex-hull problem. To see this, define the SCORE-normalized pure-profile embedding $\mathbf{g}_k \coloneqq \frac{\mathbf{U}_{S_k,2:K}^*}{U_{S_k,1}^*}$ for $k \in [K]$. Using the identity

$\mathbf{U}_i^* = \sum_{k=1}^K a_{ik}^* \mathbf{U}_{S_k}^*$, any row with at least one active feature satisfies

$$\widetilde{\mathbf{U}}_i^* = \frac{\mathbf{U}_{i,2:K}^*}{U_{i,1}^*} = \sum_{k:a_{ik}^*=1} \frac{a_{ik}^* U_{S_k,1}^*}{\sum_{\ell=1}^K a_{i\ell}^* U_{S_\ell,1}^*} \frac{\mathbf{U}_{S_k,2:K}^*}{U_{S_k,1}^*} = \sum_{k:a_{ik}^*=1} \omega_{ik} \mathbf{g}_k, \qquad \omega_{ik} = \frac{a_{ik}^* U_{S_k,1}^*}{\sum_{\ell=1}^K a_{i\ell}^* U_{S_\ell,1}^*}.$$

Under the positivity condition on the leading singular coordinate, the weights ω_{ik} are non-negative and sum to one over the active feature set. Thus, at the population level, SCORE normalization maps the vertices of the pre-projection parallelotope onto an affine simplex: pure latent profiles map exactly to the simplex vertices, while overlapping profiles map to the *face* spanned by their active pure features. When $K = 3$, subjects possessing two latent features lie on the corresponding simplex edge and subjects with all three features lie in the interior. Consequently, identifying these simplex vertices enables the recovery of the pure latent profiles, which paves the way for the estimation of the complete overlapping structure. This vertex hunting can be achieved using the Successive Projection Algorithm (SPA; Gillis and Vavasis, 2013). At the sample level, although empirical noise causes the projected data to deviate from the exact oracle vertices, we specify reasonable conditions under which the projected singular subspace remains well-separated, so that applying SPA to the empirical data successfully recovers the underlying pure latent profiles as the first step.

While previous literature on mixed membership and degree-corrected mixed membership models has similarly identified and leveraged simplex geometry for estimation (Jin, 2015; Jin et al., 2024; Ke and Wang, 2024), the role of the simplex is different in CRYSTAL. In degree-corrected mixed membership models, SCORE normalization removes node-level degree heterogeneity, and the resulting simplex coordinates are used to directly estimate continuous membership weights. In contrast, prior to projection, our overlapping clustering model admits an exact parallelotope structure, representing each subject as $\mathbf{U}_i^* = \sum_{k=1}^K a_{ik}^* \mathbf{U}_{S_k}^*$, with binary assignments $a_{ik} \in \{0, 1\}$ for all $(i, k) \in [N] \times [K]$. The barycentric weights created by SCORE are therefore not the target model parameters; they are induced by the first singular coordinate and are used only to expose the K principal edge directions of the

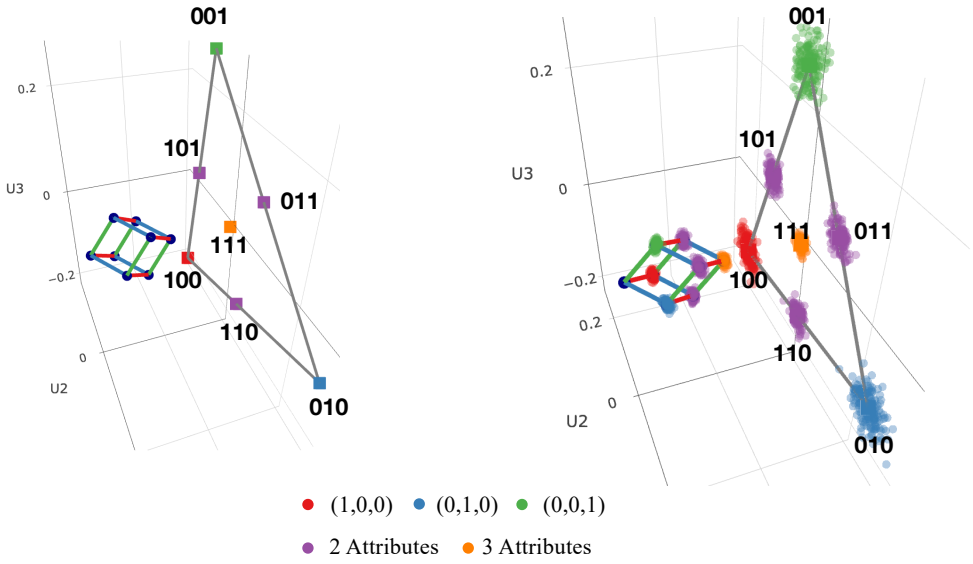

Figure 2: SCORE-normalized geometry of the singular subspace when $K = 3$. The left panel shows the population noiseless case, where pure profiles are mapped to simplex vertices and overlapping profiles to structured interior or edge points. The right panel shows the sample-based noisy case, with the corresponding noisy embeddings around the population truths.

underlying parallelotope. After these pure directions are identified, CRYSTAL returns to the unnormalized singular subspace and recovers the discrete matrix $\mathbf{A}$ through the linear reconstruction $\mathbf{U}\mathbf{U}_{\widehat{S},:}^{-1}$ followed by hard thresholding, a step that has no direct analogue in continuous mixed membership estimation.

We next describe the specific estimation procedure. For any matrix $\mathbf{M}$, let $\mathbf{M}^{\dagger}$ denote its Moore–Penrose pseudoinverse. First, we apply the top-K SVD to the data matrix $\mathbf{R}$ to obtain the top-K left singular vectors $\mathbf{U} \in \mathbb{R}^{N \times K}$, which approximate the population singular subspace $\mathbf{U}^*$. Next, we perform the SCORE normalization by dividing each row of $\mathbf{U}$ by its first coordinate, forming $\widetilde{\mathbf{U}}$, and then apply the successive projection algorithm (SPA, Gillis and Vavasis, 2013) on the rows of $\widetilde{\mathbf{U}}$ to identify a set of K indices corresponding to the K pure latent profiles. Given the estimated index set $\widehat{S} = (\widehat{S}_1, \ldots, \widehat{S}_K)$ of K single-attribute profiles, we estimate the binary latent features by first obtaining $\widetilde{\mathbf{A}} = \mathbf{U}\mathbf{U}_{\widehat{S},:}^{-1}$ and then thresholding each entry of $\widetilde{\mathbf{A}}$ at 0.5 to obtain the binary matrix $\widehat{\mathbf{A}}$. Finally, we estimate $\boldsymbol{\Theta}^*$ by least squares, equivalently, $\widehat{\boldsymbol{\Theta}}^{\top} = \widehat{\mathbf{A}}^{\dagger}\mathbf{R}$.

Algorithm 1: CRYSTAL for models without intercepts

Input: Data matrix $\mathbf{R}$, number of latent features K;
Perform top-K SVD to $\mathbf{R}$ and obtain $(\mathbf{U}, \mathbf{\Lambda}, \mathbf{V})$;
Compute the SCORE-normalized left singular vector matrix $\widetilde{\mathbf{U}} \coloneqq \mathrm{diag}(\mathbf{u}_1)^{-1}\mathbf{U}_{:,2:K} \in \mathbb{R}^{N\times(K-1)}$;
Apply the SPA to $\widetilde{\mathbf{U}}$ to obtain an estimate $\widehat{S}$ for the single-attribute indices;
Latent feature estimation: $\widetilde{\mathbf{A}} \coloneqq \mathbf{U}\mathbf{U}^{-1}_{\widehat{S},:}$;
Binarize $\widetilde{\mathbf{A}}$ to obtain $\widehat{\mathbf{A}}$ with entries $\widehat{A}_{ij} = \mathbb{1}(\widetilde{A}_{ij} > 0.5)$.
Estimate loading parameters $\mathbf{\Theta}^*$ by $\widehat{\mathbf{\Theta}}^\top \coloneqq \widehat{\mathbf{A}}^\dagger\mathbf{R}$;

3.2 Algorithm for Models with Intercepts

Compared with the cases without intercepts, the inclusion of intercepts introduces some additional complexity. When the intercept is present, the rank of the expectation matrix is $K+1$ in general. However, in some degenerate cases where every row of $\mathbf{A}^*$ is a canonical basis vector, the rank of the expectation matrix $\mathbf{R}^*$ reduces to K. Since the degree of overlap (i.e., how far the model is from the non-overlapping clustering case) is not known *a priori*, we use the following criterion: If $\sigma_K - \sigma_{K+1} \geq \frac{1}{2}\sigma_K$, then we retain the top K singular components, since this eigengap guarantees consistent estimation of the leading singular subspace by Wedin's theorem. If $\sigma_K - \sigma_{K+1} < \frac{1}{2}\sigma_K$, we retain the top $K+1$ components, again by Wedin's theorem, now using the gap between σ_{K+1} and zero.

The procedure is summarized in Algorithm 2. After computing the truncated SVD with the selected rank K_0, we first estimate the pure-intercept rows of $\mathbf{A}^*$ by examining the absolute values of the entries of the leading left singular vector $\mathbf{u}_1$. This step is reliable for two reasons: first, we impose an eigengap condition between the largest and second-largest singular values of $\mathbf{R}^*$, which guarantees the stability of the leading singular direction; second, Proposition 2 ensures that the pure-intercept rows are identifiable at the population level. Let $\widehat{S}_0 \subseteq [N]$ denote the resulting estimate of the pure-intercept index set. We then subtract the estimated intercept-profile embedding from the remaining rows of $\mathbf{U}$, yielding a pruned matrix $\mathbf{U}^{\mathsf{prune}} \in \mathbb{R}^{(N-|\widehat{S}_0|)\times K_0}$. This pruning step recenters the singular-subspace

embedding by removing the intercept contribution; after this centering, the nonzero latent-feature rows retain the same additive edge geometry as in the intercept-free case. We then apply SCORE normalization to the pruned rows, forming $\widetilde{\mathbf{U}}^{\mathsf{prune}}$, which converts the pruned edge-sum geometry into the simplex geometry used by SPA.

Regarding the choice of ψ_0 in Algorithm 2, we suggest using $\psi_0 = C\frac{\sigma\sqrt{\log(N\vee J)}}{\sigma_1}$ to identify all the pure-intercept, zero-attribute rows in $\mathbf{A}$. The rationale is that ψ_0 should be calibrated to the scale of the estimation error along the leading singular-coordinate direction. Specifically, by the singular subspace perturbation theory of Chen et al. (2021), and under suitable regularity conditions, the entrywise error $U_{i,1}\mathrm{sign}(\mathbf{U}_{:,1}^\top\mathbf{U}_{:,1}^*) - U_{i,1}^*$[1] admits an asymptotically linear approximation and is therefore asymptotically normal. Consequently, $U_{i,1}$ concentrates around its population counterpart $U_{i,1}^*$. A diagnostic illustration of this concentration phenomenon is provided in Figure S.1 of the Supplementary Material. Thus, ψ_0 may be interpreted as a data-driven estimate of the standard deviation of $U_{i,1}\mathrm{sign}(\mathbf{U}_{:,1}^\top\mathbf{U}_{:,1}^*) - U_{i,1}^*$, inflated by the logarithmic factor $C\sqrt{\log(N\vee J)}$ to ensure the desired high-probability control. This choice of threshold is in the same spirit as Bhattacharya et al. (2023), and is further justified by Remark S.3. In all simulations and real-data applications, we set $C = 2$.

4 Theoretical Guarantees for Recovery and Estimation

Next, we establish latent feature recovery guarantees as well as parameter estimation error bounds for binary latent feature models with and without intercepts in Sections 4.1 and 4.2, respectively. We will leverage the singular subspace perturbation theory to characterize the fine-grained row-wise fluctuation of $\mathbf{U}$ as well as its linear approximation form.

Before proceeding, we recap and further introduce the notations needed below. For the intercept-free model, let $\mathbf{R}^* = \mathbf{U}^*\mathbf{\Sigma}^*\mathbf{V}^{*\top}$ denote its rank-K singular value decomposition, where $\mathbf{\Sigma}^* = \mathrm{diag}(\sigma_1^*, \ldots, \sigma_K^*)$ with $\sigma_1^* \geq \cdots \geq \sigma_K^* > 0$. We define the row and column coherence parameters as $\mu_1 := \frac{N}{K}\max_{i\in[N]}\|\mathbf{U}_{i,:}^*\|_2^2$, $\mu_2 := \frac{J}{K}\max_{j\in[J]}\|\mathbf{V}_{j,:}^*\|_2^2$, so that

[1]Here, $\mathbf{U}^*$ denotes the matrix of the top $(K+1)$ left singular vectors of $\mathbf{R}^*$.

Algorithm 2: CRYSTAL for models with intercepts

Input: Data matrix $\mathbf{R}$, number of latent features K, thresholding level ψ_0;
Output: Latent feature matrix estimate $\widehat{\mathbf{A}}$, intercept estimate $\widehat{\mathbf{d}}$, and item parameter estimate $\widehat{\mathbf{\Theta}}$. ;
Compute the top $K+1$ singular values of $\mathbf{R}$ and denote them by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{K+1}$;
if $\sigma_K - \sigma_{K+1} \leq \frac{1}{2}\sigma_K$ **then**
  Set $K_0 = K+1$;
end
else
  Set $K_0 = K$;
end
Perform the top-K_0 SVD to $\mathbf{R}$ and obtain $(\mathbf{U}, \mathbf{\Lambda}, \mathbf{V})$;
$i_0 \coloneqq \arg\min_{i\in[N]} |(\mathbf{u}_1)_i|$;
$\widehat{S}_0 \coloneqq \big\{i \in [N] : |U_{i,1} - U_{i_0,1}| \leq \psi_0\big\}$;
$\mathbf{U}^{\mathsf{prune}} \coloneqq \mathbf{U}_{-\widehat{S}_0,:} - \mathbf{1}_{N-|\widehat{S}_0|} \cdot \sum_{i\in\widehat{S}_0} \mathbf{U}_{i,:}/|\widehat{S}_0|$;
$\widetilde{\mathbf{U}}^{\mathsf{prune}} = \mathrm{diag}\big((\mathbf{U}^{\mathsf{prune}})_{:,1}\big)^{-1}(\mathbf{U}^{\mathsf{prune}})_{:,2:K_0}$;
Apply the SPA to $\widetilde{\mathbf{U}}^{\mathsf{prune}}$ to obtain an estimate $\widehat{S}$ for the single-attribute pattern indices;
Latent feature estimation: $\widetilde{\mathbf{A}} \coloneqq \mathbf{U}^{\mathsf{prune}}\mathbf{U}^{\mathsf{prune}\dagger}_{\widehat{S},:}$;
Binarize $\widetilde{\mathbf{A}}$ to obtain $\check{A}_{ij} = \mathbb{1}(\widetilde{A}_{ij} > 0.5)$.
Let $\widehat{\mathbf{A}} \in \{0,1\}^{N\times K}$ with $\widehat{\mathbf{A}}_{[N]\setminus\widehat{S}_0} = \check{\mathbf{A}}$, and $\widehat{\mathbf{A}}_{\widehat{S}_0} = \mathbf{0}_{|\widehat{S}_0|\times K}$.
We define $\widehat{\mathbf{A}}^{\mathsf{app}} = [\mathbf{1}_N \mid \widehat{\mathbf{A}}]$. $\begin{pmatrix}\widehat{\mathbf{d}}^\top \\ \widehat{\mathbf{\Theta}}^\top\end{pmatrix} \coloneqq (\widehat{\mathbf{A}}^{\mathsf{app}})^\dagger\mathbf{R}$;

$\max_i \|\mathbf{U}^*_{i,:}\|_2 \leq \sqrt{\mu_1 K/N}$ and $\max_j \|\mathbf{V}^*_{j,:}\|_2 \leq \sqrt{\mu_2 K/J}$. For the model with intercept, the leading singular subspace is defined slightly differently: if $\sigma^*_{K+1} \neq 0$, then $(\mathbf{U}^*, \mathbf{\Sigma}^*, \mathbf{V}^*)$ is taken to be the top-$(K+1)$ SVD of $\mathbf{R}^*$; otherwise, it is taken to be the top-K SVD of $\mathbf{R}^*$. The incoherence degrees μ_1 and μ_2 are then defined through $\mathbf{U}^*$ and $\mathbf{V}^*$, respectively.

We use the following notation throughout the theoretical statements. For a matrix $\mathbf{M}$, let $\sigma_k(\mathbf{M})$ denote its kth largest singular value, let $\sigma_{\min}(\mathbf{M})$ denote its smallest nonzero singular value, and define $\kappa(\mathbf{M}) = \sigma_1(\mathbf{M})/\sigma_{\min}(\mathbf{M})$. In particular, for the population signal matrix $\mathbf{R}^*$, write $\kappa^* = \sigma^*_1/\sigma^*_K$. We use $\|\mathbf{M}\|$ for the spectral norm, $\|\mathbf{M}\|_\infty$ for the entrywise maximum norm, and $\|\mathbf{M}\|_{2,\infty} = \max_i \|\mathbf{M}_{i,:}\|_2$ for the row-wise two-to-infinity norm. Let

$\mathsf{perm}([\mathsf{K}])$ denote the set of $K \times K$ permutation matrices. The upper noise-variance scale is denoted by $\sigma^2 = \max_{i,j} \mathrm{Var}(E_{i,j})$. Finally, $a(N,J) \lesssim b(N,J)$ (resp. $a(N,J) \gtrsim b(N,J)$) means $a(N,J) \leq Cb(N,J)$ (resp. $C_n a(N,J) \geq b(N,J)$) for a universal constant $C > 0$, $a(N,J) \asymp b(N,J)$ means both $a(N,J) \lesssim b(N,J)$ and $b(N,J) \lesssim a(N,J)$, and $a(N,J) \ll b(N,J)$ (resp. $a(N,J) \gg b(N,J)$) means there exists some sufficiently small (resp. large) constant c such that $a(N,J) \leq cb(N,J)$ (resp. $a(N,J) \geq cb(N,J)$) for all sufficiently large N and J.

We impose the following assumptions, common to both models with and without intercepts, to control the noise level, guarantee sufficient signal strength, and ensure a nontrivial eigengap between the largest and second-largest singular values of the population matrix $\mathbf{R}^*$.

Assumption 1. *We assume that* $K \ll \min\{N, J\}$*, and*

$$\sigma_K(\mathbf{A}^*)\sigma_K(\mathbf{\Theta}^*) \gg \sigma\mu_1(\kappa^*)^2 K^{\frac{5}{2}}\kappa^2(\mathbf{A}^*)\sqrt{(N \vee J)\log(N \vee J)}.$$

Assumption 2. *We assume that* $\triangle := \sigma_1(\mathbf{R}^*) - \sigma_2(\mathbf{R}^*) \gg \sigma\mu_1 K^{\frac{5}{2}}\kappa^2(\mathbf{A}^*)\sqrt{N \vee J}$..

Assumption 3. *Assume that the noise matrix* $\mathbf{E}$ *satisfies (a)* $\mathbf{E}$ *is entrywise independent, (b) Either* $\|E_{i,j}\|_\infty \leq B$ *for all* $i \in [N]$, $j \in [J]$, *or there exists a random matrix* $\mathbf{E}' = (E'_{i,j}) \in \mathbb{R}^{N \times J}$, *such that for any* $i \in [N], j \in [J]$, *it holds that* $\left\|E'_{i,j}\right\|_\infty \leq B$, $\mathbb{E}[E'_{i,j}] = 0$, *and* $\Pr(E_{i,j} = E'_{i,j}) \geq 1 - O((N \vee J)^{-22})$. *Also, we assume that* $B\log^2(N \vee J)\max\left\{\sqrt{\frac{\mu_1 K}{N}}, \sqrt{\frac{\mu_2 K}{J}}\right\}\max\left\{1, \frac{J^2}{N^2}\right\} \lesssim \sigma$ *and* $(\log(N \vee J))^4 \ll N \wedge J$.

Assumption 3 covers a broad class of noise models. In particular, assuming $\frac{J}{N}, \mu_1, \mu_2 = O(1)$, it includes *all sub-Gaussian and sub-exponential distributions* where the corresponding norms are within a constant factor of their standard deviations. It also encompasses sparse Bernoulli distributions with means at least the magnitude of $\frac{1}{N \wedge J}$ up to logarithmic factors.

4.1 Models without Intercepts

We first consider the intercept-free model. To ensure statistical stability, we impose the following assumption, which provides a quantitative version of the condition in Proposition 1

and appears as a standard requirement in the analysis of mixed membership models (Jin, 2015; Jin et al., 2024).

Assumption 4. *Suppose the top eigenvector of* $\mathbf{\Theta}^{*\top}\mathbf{\Theta}^{*}\mathbf{A}^{*\top}\mathbf{A}^{*}$ *is positive, or equivalently, the entries of the leading left singular vector satisfy* $\min_{i\in[N]} U^{*}_{i,1} > 0$. *Further, suppose the ratio among the entries of* $\mathbf{\Theta}^{*\top}\mathbf{\Theta}^{*}\mathbf{A}^{*\top}\mathbf{A}^{*}$*'s leading eigenvector is bounded by a constant.*

In Section S.3.1 of the Supplementary Material, we discuss scenarios where Assumption 4 holds. A simulation diagnostic for this condition is reported in Figure S.2 of the Supplementary Material.

Theorem 1. *Suppose that the model satisfies the conditions in Proposition 1 and Assumptions 1, 2, 3, and 4 hold. Then,* $\mathbb{P}\big[\exists \mathbf{\Pi} \in \mathsf{perm}([K])\ s.t.\ \widehat{\mathbf{A}}\mathbf{\Pi} = \mathbf{A}^{*}\big] = 1 - O((N \vee J)^{-10})$.

We further establish the estimation error bounds for $\mathbf{\Theta}^{*}$.

Theorem 2 ($\mathbf{\Theta}^{*}$ Estimation Error)**.** *Under the assumptions in Theorem 1, with probability at least* $1 - O((N \vee J)^{-10})$, *we have*

$$\min_{\mathbf{\Pi}\in\mathsf{perm}([K])} \big\|\widehat{\mathbf{\Theta}}\mathbf{\Pi} - \mathbf{\Theta}^{*}\big\| = \big\|\mathbf{E}^{\top}(\mathbf{A}^{*})^{\dagger\top}\big\| \lesssim \frac{\sigma\sqrt{J}}{\sigma_{\min}(\mathbf{A}^{*})},$$
$$\min_{\mathbf{\Pi}\in\mathsf{perm}([K])} \big\|\widehat{\mathbf{\Theta}}\mathbf{\Pi} - \mathbf{\Theta}^{*}\big\|_{\infty} = \big\|\mathbf{E}^{\top}(\mathbf{A}^{*})^{\dagger\top}\big\|_{\infty} \lesssim \frac{\sigma\sqrt{K\log(N \vee J)}}{\sigma_{K}(\mathbf{A}^{*})}.$$

In comparison, existing likelihood-based studies such as Gu and Xu (2023) mainly establish average consistency for the parameter estimator, which does not directly imply entrywise accuracy of $\mathbf{\Theta}^{*}$. By contrast, our ℓ_{∞} bound yields entrywise consistency, a sharper form of control that is useful for plug-in estimation of downstream functionals such as covariance-type quantities, paving the way for performing statistical inference.

4.2 Models with Intercepts

To handle models with intercepts, we augment the parameters by introducing their appended versions, denoted by $\mathbf{A}^{*,\mathsf{app}}$ and $\mathbf{\Theta}^{*,\mathsf{app}}$, respectively:

$$\mathbf{A}^{*,\mathsf{app}} = (\mathbf{1}_{N}, \mathbf{A}^{*}), \qquad \mathbf{\Theta}^{*,\mathsf{app}} = (\mathbf{d}^{*}, \mathbf{\Theta}^{*}).$$

To align with the identifiability condition in Proposition 2, we impose the following assumption analogous to Assumption 4.

Assumption 5. *We assume that the top eigenvector of* $\boldsymbol{\Theta}^{*,\mathsf{app}\top}\boldsymbol{\Theta}^{*,\mathsf{app}}\mathbf{A}^{*,\mathsf{app}\top}\mathbf{A}^{*,\mathsf{app}}$ *is positive, or equivalently, the entries of the leading left singular vector of* $\mathbf{R}^*$ *satisfy* $\min_{i\in[N]} U^*_{i,1} > 0$. *Further, we assume that the ratio between the largest and smallest entries of this leading eigenvector is bounded by a constant.*

Denote $S_0^* := \{i \in [N] : \mathbf{A}^*_{i,:} = \mathbf{0}\}$. We have the following recovery and estimation guarantees for Algorithm 2.

Theorem 3. *For Algorithm 2, suppose that the conditions in Proposition 2 are satisfied, and Assumptions 1, 2, 3, and 5 hold and* S_0^* *is non-empty. Then for a suitably chosen threshold* ψ_0, *we have* $\mathbb{P}\big[\exists \boldsymbol{\Pi} \in \mathsf{perm}([K])\ s.t.\ \widehat{\mathbf{A}}\boldsymbol{\Pi} = \mathbf{A}^*\big] = 1 - O((N \vee J)^{-10})$.

Theorem 4. *Under the assumptions in Theorem 3, with probability at least* $1 - O((N \vee J)^{-10})$*:*

$$\left\|\widehat{\mathbf{d}} - \mathbf{d}^*\right\|_2 \lesssim \frac{\sigma\sqrt{J}}{\sqrt{|S_0^*|}}, \qquad \left\|\widehat{\mathbf{d}} - \mathbf{d}^*\right\|_\infty \lesssim \frac{\sigma\sqrt{|S_0^*|\log(N\vee J)} + B\log(N\vee J)}{|S_0^*|},$$

$$\min_{\boldsymbol{\Pi}\in\mathsf{perm}([K])} \left\|\widehat{\boldsymbol{\Theta}}\boldsymbol{\Pi} - \boldsymbol{\Theta}^*\right\| \lesssim \frac{\sigma\sqrt{KNJ}}{\sqrt{|S_0^*|}\sigma_K(\mathbf{A}^*)^2} + \frac{\sigma\sqrt{J}}{\sigma_K(\mathbf{A}^*)},$$

$$\min_{\boldsymbol{\Pi}\in\mathsf{perm}([K])} \left\|\widehat{\boldsymbol{\Theta}}\boldsymbol{\Pi} - \boldsymbol{\Theta}^*\right\|_{2,\infty} \lesssim \frac{\big(\sigma\sqrt{|S_0^*|\log(N\vee J)} + B\log(N\vee J)\big)\sqrt{NK}}{|S_0^*|\sigma_K(\mathbf{A}^*)^2} + \frac{\sigma\sqrt{K\log(N\vee J)}}{\sigma_K(\mathbf{A}^*)}.$$

The above rates separate two sources of error. The bounds for $\widehat{\mathbf{d}}$ improve with $|S_0^*|$ through averaging over the estimated pure-intercept rows. The bounds for $\widehat{\boldsymbol{\Theta}}$ depend on both $|S_0^*|$ and $\sigma_K(\mathbf{A}^*)$: the first terms propagate intercept-estimation error, while the second terms arise from projecting the noise matrix onto $\mathbf{A}^*$. These bounds imply entrywise consistency of the continuous-parameter estimators in suitably high-dimensional regimes.

5 Statistical Inference

This section develops statistical inference procedures, along with accompanying theoretical guarantees, for the population parameters $\mathbf{d}^*$ and $\mathbf{\Theta}^*$. As discussed in Section 4, our assumptions already ensure exact recovery of $\mathbf{A}^*$ with high probability, so separate inference for $\mathbf{A}^*$ is unnecessary.

To perform statistical inference for $\mathbf{\Theta}^*$, we estimate the limiting covariance of $\widehat{\mathbf{\Theta}}_{j,:}$ from Algorithm 1 or Algorithm 2. We focus on settings in which the covariance of each entry $E_{i,j}$ can be expressed as a function g of its corresponding expectation. This includes Bernoulli and Poisson observations, our motivating settings. We introduce the following estimators.

1. *Intercept-free case*: We use $\mathsf{diag}\big(g(\widehat{R}_{i,j})\big)_{i\in[N]}$ as a plug-in estimator for $\mathrm{Cov}(\mathbf{E}_{\cdot,j}) = \mathrm{diag}(\mathrm{Var}(E_{i,j}))_{i\in[N]}$, and $\widehat{\mathbf{A}}$ as an estimator of $\mathbf{A}^*$. Motivated by the linear expansion of $\widehat{\mathbf{\Theta}}_{j,:} - \mathbf{\Theta}^*_{j,:} \approx \mathbf{E}_{\cdot,j}^\top \mathbf{A}^{*\dagger}$, we define $\widehat{\mathbf{\Sigma}}_j \coloneqq \widehat{\mathbf{A}}^\dagger \mathsf{diag}\big(g(\widehat{R}_{i,j})\big)_{i\in[N]} \widehat{\mathbf{A}}^{\dagger\top}$.

2. *Case with intercept*: In the presence of an intercept, the covariance structure includes the estimation error of the intercept term. To account for this additional contribution, we propose the estimator

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_j \coloneqq &(\widehat{\mathbf{A}}_{-\widehat{S}_0,:})^\dagger \mathsf{diag}\big(g(\widehat{R}_{i,j})\big)_{i\in[N]\setminus\widehat{S}_0} (\widehat{\mathbf{A}}_{-\widehat{S}_0,:})^{\dagger\top} \\ &+ \frac{1}{|\widehat{S}_0|^2} (\widehat{\mathbf{A}}_{-\widehat{S}_0,:})^\dagger \mathbf{1}_{N-|\widehat{S}_0|} \mathbf{1}_{|\widehat{S}_0|}^\top \mathsf{diag}\big(g(\widehat{R}_{i,j})\big)_{i\in\widehat{S}_0} \mathbf{1}_{|\widehat{S}_0|} \mathbf{1}_{N-|\widehat{S}_0|}^\top (\widehat{\mathbf{A}}_{-\widehat{S}_0,:})^{\dagger\top},\end{aligned}$$

   where $\widehat{\mathbf{R}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ denotes the rank-$K+1$ SVD approximation of the data matrix $\mathbf{R}$.

The next theorem shows the asymptotic normality of the item parameters estimator $\widehat{\mathbf{\Theta}}$.

Theorem 5. *Suppose the assumptions of Theorem 1 hold for the intercept-free model, or the assumptions of Theorem 3 and $\sigma^*_{K+1}/\sigma^*_K \gtrsim 1$ hold for the model with an intercept. Additionally, assume that the variance of each entry $R_{i,j}$ satisfies $\mathrm{Var}(R_{i,j}) = g(R^*_{i,j})$, where g is a fixed Lipschitz function. Then with probability at least $1 - O((N \vee J)^{-10})$, there exists*

a permutation matrix $\widehat{\mathbf{\Pi}}$ *such that*

$$\sup_{C\in\mathfrak{C}^K}\left|\mathbb{P}\big[\big(\widehat{\mathbf{\Theta}}_{j,:}\widehat{\mathbf{\Pi}}-\mathbf{\Theta}^*_{j,:}\big)\in C\big]-\mathbb{P}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_K)}[\mathbf{\Sigma}_j^{\frac{1}{2}}\mathbf{x}\in C]\right|\lesssim\frac{K^{\frac{1}{4}}BN^{-\frac{1}{2}}}{\min_{i\in[N],j\in[J]}\ Var(E_{i,j})^{\frac{1}{2}}}+(N\vee J)^{-10},$$

where $\mathbf{\Sigma}_j\coloneqq\mathbf{A}^{*\dagger}\mathrm{diag}\big(g(R^*_{i,j})\big)_{i\in[N]}\mathbf{A}^{*\dagger^\top}$ *for the intercept-free model and*

$$\begin{aligned}\mathbf{\Sigma}_j\coloneqq&(\mathbf{A}^*_{-S^*_0,:})^\dagger\mathsf{diag}\big(g(R^*_{i,j})\big)_{i\in[N]\setminus S^*_0}(\mathbf{A}^*_{-S^*_0,:})^{\dagger\top}\\&+\frac{1}{|S^*_0|^2}(\mathbf{A}^*_{-S^*_0,:})^\dagger\mathbf{1}_{N-|S^*_0|}\mathbf{1}^\top_{|S^*_0|}\mathsf{diag}\big(g(R^*_{i,j})\big)_{i\in S^*_0}\mathbf{1}_{|S^*_0|}\mathbf{1}^\top_{N-|S^*_0|}(\mathbf{A}^*_{-S^*_0,:})^{\dagger\top}\end{aligned}$$

for models with intercepts. Here, $\mathfrak{C}^K$ *denotes the collection of all Borel convex sets in* $\mathbb{R}^K$*.*

If, in addition, the noise variance satisfies

$$\min_{i\in[N],j\in[J]}\mathrm{Var}(E_{i,j})\gg\max\Big\{\frac{K^{\frac{7}{2}}\mu_1B^2}{N},\frac{\kappa^*\kappa^2(\mathbf{A}^*)K^2\log(N\vee J)\sqrt{\mu_1\mu_2}}{\sqrt{N\wedge J}}\Big\},\quad\textit{(without intercept)},$$

$$\min_{i\in[N],j\in[J]}\mathrm{Var}(E_{i,j})\gg\max\Big\{\frac{K^{\frac{7}{2}}\mu_1B^2}{|S^*_0|},\frac{\kappa^*\kappa(\mathbf{A}^*)^2K^2\log(N\vee J)\sqrt{\mu_1\mu_2}N}{\sqrt{N\wedge J}|S_0|}\Big\},\ \textit{(with intercept)}.$$

Then we have for the statistic normalized by our proposed covariance estimators that

$$\sup_{\mathbf{z}\in\mathbb{R}^K}\left|\mathbb{P}\big[\big(\widehat{\mathbf{\Theta}}_{j,:}\widehat{\mathbf{\Pi}}-\mathbf{\Theta}^*_{j,:}\big)\widehat{\mathbf{\Sigma}}_j^{-\frac{1}{2}}\le\mathbf{z}\big]-\Phi(\mathbf{z})\right|=o(1).$$

We next explain the conditions under which Theorem 5 ensures asymptotically valid type-I error control. Beyond the exact-recovery signal constraints, the distributional convergence requires a mild lower bound on the minimum variance of entries. Assuming κ^*, $\kappa(\mathbf{A}^*)$, K, μ_1, μ_2, $N/|S_0|$, and $|\log(N/J)|$ are all $O(1)$, asymptotic normality holds for both settings if $B\le C_1(\log(N\vee J))^{C_2}$ and $\min_{i\in[N],j\in[J]}\mathrm{Var}(E_{i,j})\ge C_3$ for some constants $C_1,C_2,C_3>0$. This, in turn, guarantees valid type-I error control.

6 Simulation Studies

In this section, we conduct extensive simulations to assess the performance of CRYSTAL with or without intercepts under a wide variety of circumstances.

6.1 Simulation to Assess Parameter Estimation

First, we evaluate CRYSTAL against the standard marginal maximum likelihood estimation (MMLE) via the EM algorithm (de la Torre, 2011) in the intercept-free case. We utilized the MMLE implementation provided in the `R` package `GDINA` (Ma and de la Torre, 2020). In this simulation, we implement the CRYSTAL algorithm under the assumption of a sparsity structure $\mathbf{Q}$ governing the item parameters $\boldsymbol{\Theta}$; that is, if $q_{jk} = 0$, $\theta_{jk} = 0$. Particularly within the framework of Cognitive Diagnostic Models (CDMs; Templin and Henson, 2006), imposing such a sparse structure on $\boldsymbol{\Theta}$ is standard practice, motivated by both empirical interpretability and theoretical identifiability. Empirically, individual items in educational assessments are typically designed to evaluate only a specific subset of latent attributes. Theoretically, an underlying sparse $\mathbf{Q}$-matrix is fundamental to establishing the *population identifiability* of CDMs (Xu and Shang, 2018). Although our theoretical framework circumvents the need for item-level sparsity by introducing mild identifiability conditions on the overlapping clustering structure (Propositions 1–2), we incorporate it here to facilitate a direct comparison with the broader CDM literature. Furthermore, by leveraging the asymptotic distribution of the item parameters derived in Section 5, CRYSTAL can estimate this underlying sparsity structure. This capability directly connects our methodology to the extensive psychometric literature on $\mathbf{Q}$-matrix estimation (Liu et al., 2012; Chen et al., 2015).

It is important to clarify that even though the two estimation methods under comparison are developed for the same model, the Additive Cognitive Diagnostic Model (ACDM) proposed by de la Torre (2011), the MMLE algorithm in the `GDINA` package requires the correct $\mathbf{Q}$-matrix as an input. The MMLE algorithm is supplied with the oracle $\mathbf{Q}$, whereas CRYSTAL estimates it together with other parameters. Consequently, CRYSTAL faces a significantly more challenging estimation task than the MMLE.

The simulation setting is as follows. To ensure identifiability, the binary features matrix $\mathbf{A}$ contains five identity matrices. The remaining entries are generated from a Bernoulli

distribution with $p = 0.5$. To prevent null attribute profiles, any row generated as a zero vector is replaced by a standard basis vector $\{\mathbf{e}_k\}_{k=1}^K$ with uniform probability $1/K$. The $\mathbf{Q}$-matrix is constructed to represent varying item complexities. Specifically, $J/2$ items load on a single attribute, while $J/4$ items load on two and three attributes, respectively. Based on $\mathbf{Q}$, the item parameters in $\boldsymbol{\Theta}$ are generated so that the nonzero entries in each row are equal and sum to 0.8. All simulations are conducted with $N = 3000$ and $J = 1000$, with K varying across $\{10, 12, 14, 16\}$, corresponding to challenging estimation scenarios. Each simulation setting contains 100 independent replicates. This simulation setting matches standard practices in previous literature (Gu and Xu, 2023).

Table 1 summarizes the simulation results. The simulation setting inherently favors MMLE in two aspects. First, MMLE is provided with the oracle $\mathbf{Q}$, bypassing the challenge of estimating the Q-matrix. Second, the computational complexity of MMLE limits its application to smaller N, J, and K. MMLE becomes computationally intractable as the data dimensions or latent dimension grow, as indicated in the simulations when $K \in \{20, 25, 30\}$.

In terms of computation efficiency, CRYSTAL is significantly faster than the MMLE algorithm across all settings. While the computational complexity of MMLE increases exponentially with K, CRYSTAL scales linearly. In terms of recovering the binary latent features, both methods achieve exact recovery across every simulation replicate. Even in regimes where MMLE fails due to the exponential computational complexity with increasing K, CRYSTAL maintains low estimation error. Regarding the estimation error of item parameters, CRYSTAL performs slightly worse than MMLE. This is primarily attributed to MMLE's access to the oracle Q-matrix. For CRYSTAL, small errors in the estimated sparsity structure propagate to the item parameter estimation. However, the overall estimation accuracy of CRYSTAL remains highly competitive with the oracle-aided MMLE.

The simulations are designed to be compatible with the signal-strength conditions in Section 4, and the exact recovery results are consistent with the theoretical recovery guarantees

under these simulated regimes.

	K	10	12	14	16	20	25	30
Time	CRYSTAL	1.63	1.91	2.18	2.20	20.47	18.98	18.67
	MMLE	17.2	19.2	42.7	147.	–	–	–
A	CRYSTAL	0	0	0	0	5e-7	4.13e-6	2.4e-5
	MMLE	0	0	0	0	–	–	–
Theta/10^{-3}	CRYSTAL	2.12	1.81	1.60	1.47	1.30	1.19	1.17
	MMLE	1.87	1.57	1.35	1.18	–	–	–

Table 1: Comparison of MMLE and CRYSTAL across different values of K. Metrics for $\mathbf{A}$ and $\mathbf{\Theta}$ are mean absolute errors, and entries are averaged over 100 independent replicates. Dashes indicate that MMLE was not run; runtime entries should be interpreted together with the problem dimensions for each setting.

To demonstrate the performance of CRYSTAL beyond the computational limits of MMLE, we conduct additional simulations with larger latent dimensions $K \in \{20, 25, 30\}$ for both Bernoulli and Poisson data.

For the Bernoulli case, $\mathbf{A}$, $\mathbf{Q}$, and $\mathbf{\Theta}$ are generated as in the previous simulation. We vary the ratio $N/J \in \{1/5, 1, 5\}$, with $\min\{N, J\}$ ranging across $\{500, 1000, 1500, 2000\}$. Representative results for the balanced case $N = J$ are presented in Figure 3. Additional aspect-ratio settings and diagnostics on the diversity of generated overlapping structures are reported in Section S.7 of the Supplementary Material.

For the Poisson case, we evaluate the performance of CRYSTAL under both sparse and non-sparse specifications of the item parameters $\mathbf{\Theta}$. Imposing a sparse structure on $\mathbf{\Theta}$ is a standard approach to enhance interpretability when modeling sparse responses, like in environmental studies (Zhou et al., 2025). In the sparse setting, $\mathbf{A}$ and $\mathbf{Q}$ follow the generation process for the Bernoulli case, while the nonzero entries of $\mathbf{\Theta}$ are set to $\theta_{jk} = 1.5 + \varepsilon_{jk}$, with $\varepsilon_{jk} \sim \text{Exp}(1)$. In the dense setting, every entry of $\mathbf{\Theta}$ is independently drawn from Exp(1). The (N, J) and replication counts are the same as in the Bernoulli case.

As N and J increase, the estimation error empirically decreases, consistent with the the-

oretical pattern in Section 4. Even with limited sample sizes and large K, the error remains well-controlled. Notably, the estimation error for $\mathbf{\Theta}$ actually decreases as K increases. This counterintuitive trend is attributable to sparsity: as K grows, the proportion of nonzero entries in $\mathbf{\Theta}$ diminishes. Consequently, once N and J are sufficient to recover the sparsity structure accurately, the estimation error drops more rapidly for higher K. Furthermore, at a moderate sample size of $N = 2000$, CRYSTAL achieves exact recovery in nearly all scenarios, even for the challenging case of $K = 30$. Additional Poisson sparse and dense loading results are reported in Section S.7 of the Supplementary Material.

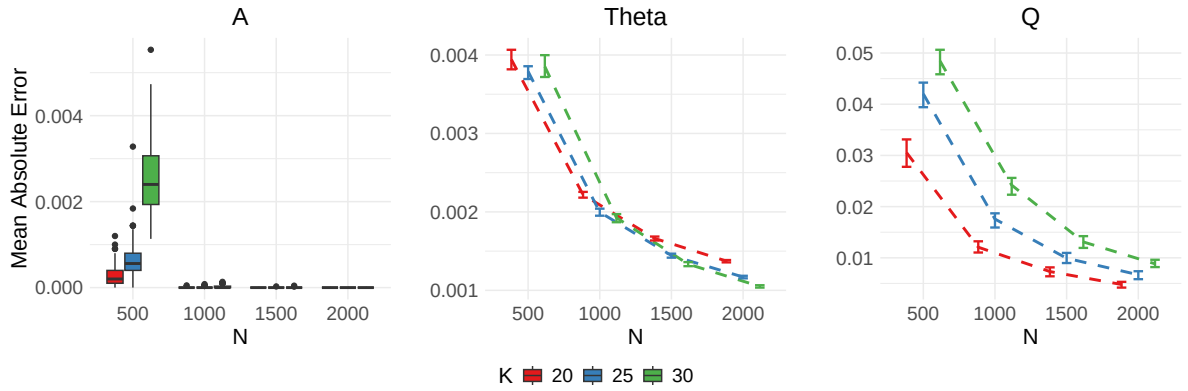

Figure 3: Bernoulli simulations with sparse loadings and $N = J$. Each panel reports mean absolute error versus N; curves denote K, and error bars mark the 25th and 75th percentiles across replicates.

Figure 4 presents simulation results for the Poisson case with the intercept. The intercept $\mathbf{d}$ is fixed at 0.5 for all entries. The attribute matrix $\mathbf{A}$ is constructed such that $\lceil N/(2(K+1))\rceil$ rows are zero vectors, $\lceil N/(3K)\rceil$ rows of identity matrices, and the remaining entries are generated independently from a Bernoulli(0.5) distribution. The matrices $\mathbf{Q}$ and $\mathbf{\Theta}$ follow the same generation process as the sparse Poisson case without the intercept. We set $N = J$ ranging from $\{500, 1000, 1500, 2000\}$, with K varying among $\{10, 12, 15\}$. Each experiment condition is replicated 100 times.

The estimation error for all parameters decreases as the sample size N increases. This trend is consistent with the theoretical pattern, particularly given that the absolute number

of full-zero entries in the overlapping structure increases with N. Notably, at $N = 2000$, CRYSTAL achieves exact recovery in all simulated cases.

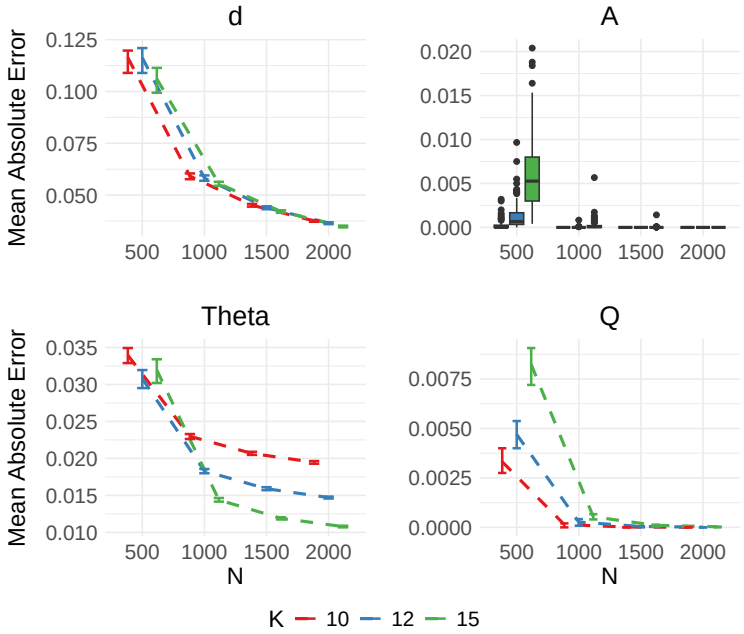

Figure 4: Poisson simulations with intercept and $N = J$. Each panel reports mean absolute error versus N; curves denote K, and error bars mark the 25th and 75th percentiles across replicates.

6.2 Simulation to Assess Statistical Inference

We next present simulations to assess our inference results. We conduct simulations for both Bernoulli and Poisson data. $\mathbf{A}$, $\mathbf{Q}$, and $\boldsymbol{\Theta}$ follow the same generation protocols as the sparse case in Section 6.1. The nonzero entries in the target testing row of $\boldsymbol{\Theta}$ are set to 0.02.

We test the null hypothesis that specific entries in the target row are zero, that is $H_0 : \theta_{j,\mathcal{I}} = \mathbf{0}$, and $H_1 : \theta_{j,\mathcal{I}} \neq \mathbf{0}$, where $\mathcal{I} \subseteq [K]$. In the context of CDMs, this is equivalent to testing whether a particular attribute is required for a given item. Further, we evaluate the power of such tests under two scenarios. For both tests, we set the null hypothesis to be $H_0 : \boldsymbol{\theta}_{j,1:2} = \mathbf{0}_2$. The alternative hypotheses for the two cases are respectively $H_1 : \boldsymbol{\theta}_{j,1:2} = (0.02, 0.02)$, and $H_1 : \boldsymbol{\theta}_{j,1:2} = (0.02, 0)$. For all simulations, we set $K = 3$ and vary $N = J$ across $\{500, 1000, 2000, 4000\}$. The significance level is $\alpha = 0.05$. In each simulation setting, we perform 1000 independent replicates and calculate the empirical rejection rates.

The average rejection rates for Bernoulli and Poisson data are in Tables 2 and 3, respectively. For the Bernoulli case, the Type I error rate for single-entry tests is well-controlled even at $N = 500$. Joint tests for multiple entries are more challenging due to the estimation of the covariance matrix. However, as N exceeds 2000, the rejection rate converges to 0.05. Furthermore, as N increases, the power of both single and joint tests progressively approaches 1. The Poisson case benefits from a stronger signal-to-noise ratio, allowing for effective Type I error control at $N = 500$ for both single and multiple entry tests.

Setting	H_0	H_1	Sample Size (N)			
			500	1000	2000	4000
H_0 is true	0	$\neq 0$	0.059	0.042	0.055	0.044
	$(0,0)$	$\neq (0,0)$	0.090	0.071	0.046	0.056
H_1 is true	$(0,0)$	$(0.02, 0.02)$	0.677	0.960	1.000	1.000
	$(0,0)$	$(0.02, 0)$	0.332	0.661	0.977	1.000

Table 2: Average rejection rates over 1000 independent replicates for Bernoulli data at significance level $\alpha = 0.05$. The top half reports type I error under the null, and the bottom half reports empirical power under the listed alternatives.

H_0	H_1	Sample Size (N)			
		500	1000	2000	4000
0	$\neq 0$	0.035	0.055	0.059	0.045
$(0,0)$	$\neq (0,0)$	0.048	0.049	0.059	0.059

Table 3: Average rejection rate of the null hypothesis over 1000 simulation replicates for Poisson data. The significance level is $\alpha = 0.05$.

Additional Q-Q diagnostics for the Bernoulli and Poisson inference simulations are in S.7 of the Supplement and are consistent with the rejection-rate summaries in Tables 2 and 3.

7 Real Data Applications

7.1 Cognitive Diagnosis from Educational Assessment Data

In this section, we apply the CRYSTAL method to the Trends in International Mathematics and Science Study (TIMSS) 2011 dataset, available from the TIMSS International Database. The dataset assesses 4th and 8th grade students across a variety of countries, recording their

responses to mathematics and science items. For our analysis, we select a subset of 8th grade students from England, USA, Finland, Israel, and Singapore, focusing on a subset of 26 mathematics questions. Responses from students are binary, with 1 denoting a correct response and 0 denoting an incorrect response. After removing observations with missing values, we obtain a final sample with dimensions $(N, J) = (2070, 26)$.

To enhance robustness in real-world data, we introduce modifications detailed in Algorithm S.1. In this adapted algorithm, we utilize the SCORE-normalized singular vectors to obtain $\widetilde{\mathbf{A}}^{\mathsf{mixed}}$. At the population level, each row of this matrix lies within a K-dimensional unit simplex. To enforce this geometric constraint, we project every row of $\widetilde{\mathbf{A}}^{\mathsf{mixed}}$ onto the unit simplex. The algorithm for such projection is well-studied and computationally efficient. Specifically, we implement the sorting-based Algorithm 1 from Condat (2016). Following the projection, we recover the final latent feature matrix by rescaling the estimates with the first singular vector. All other steps remain identical to Algorithm 2, and the full algorithm is deferred to Algorithm S.1 of the Supplementary Material.

The above modification is motivated by possible substantial noise in empirical studies. Under significant noise, empirical data may lie outside of the population parallelotope and its corresponding projected simplex. Projecting these out-of-bound points back onto the unit simplex regularizes the empirical embedding. Conceptually, this projection relates to the pruning procedure introduced by Mao et al. (2021), but it constrains all points to reside within the simplex boundaries rather than discarding outliers with sparse neighborhoods.

We leverage the reference Q-matrix provided by TIMSS in the dataset for estimation, which is constructed by the domain knowledge measured by each question item. First, we estimate the intercept and the binary latent feature matrix via Algorithm S.1. Subsequently, given the sparsity pattern Q-matrix structure, we estimate the item parameters by independently performing regressions for each item, as described in Algorithm S.2 of the Supplement. Algorithm 2 can be viewed as the fully dense special case $\mathbf{Q} = \mathbf{1}_{J\times K}$.

Figure 5 shows the estimated item-parameter heatmap. The fitted pattern is broadly

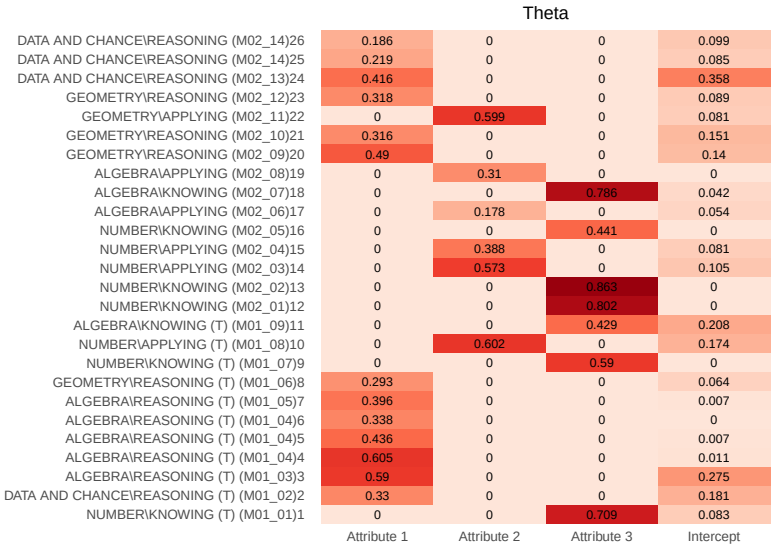

Figure 5: Estimated TIMSS item-parameter $\widehat{\boldsymbol{\Theta}}$ heatmap. Each entry reports the fitted item-response loading (item parameter) before calibration, with darker colors indicating larger fitted values. The estimated sparsity patterns indicate that we can interpret latent skill attribute 1 as "Reasoning"; attribute 2 as "Applying", and attribute 3 as "Knowing".

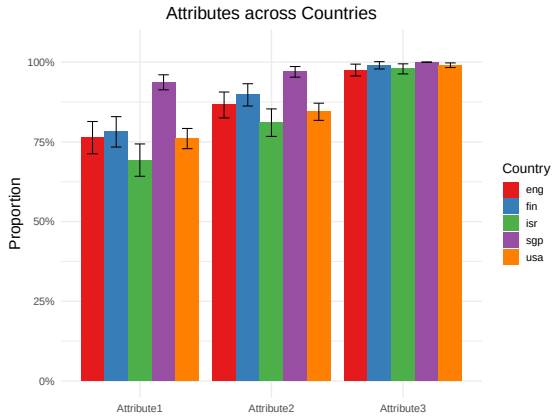

Figure 6: Proportion of participants possessing each latent attribute, stratified by country. Error bars represent ± 1.96 standard errors. `eng` stands for England, `fin` stands for Finland, `lsr` stands for Israel, `sgp` stands for Singapore, and `usa` stands for USA.

consistent with the TIMSS cognitive-domain labels used to name the latent attributes as *Reasoning*, *Applying*, and *Knowing*. Under the fitted model, the intercept represents the baseline probability of answering an item correctly for a student who possesses none of the measured latent abilities. The larger fitted loadings for Knowing are consistent with its role as a basic competency relative to Reasoning and Applying.

Figure 6 summarizes the fitted attribute proportions by country. Each plotted value is the country-specific average of the corresponding column in the estimated binary latent-feature submatrix. Under the fitted model, the USA, England, and Finland show broadly similar attribute possessions. Israel has lower fitted proportions for Reasoning and Applying, whereas Singapore has the highest fitted proportions across all three skill attributes. Since all fitted proportions exceed 50%, lower proportions correspond to greater variance in the binary latent attributes, suggesting greater heterogeneity among students. This pattern is consistent with official findings from the PISA 2022 technical report, which report relatively large between-student variation in Israel among developed economies (OECD, 2023).

7.2 Overlapping Habitat Discovery from Ecology Data

Next, we demonstrate CRYSTAL in an ecology dataset. Birds serve as an important instrument for monitoring environmental changes. Here, we analyze a Scandinavian bird dataset (Piirainen et al., 2023), a widely used benchmark in environmental research (Ovaskainen and Abrego, 2020). While this problem has traditionally been approached using continuous latent variable models (Norberg et al., 2019), recent work has adopted discrete latent variable models with overlapping clustering structures to improve interpretability while maintaining model expressivity (Scherting and Dunson, 2024; Zhou et al., 2025).

The original data record species abundance across various locations annually from 1975 to 2016. To avoid temporal dependence, we select data exclusively from the year 2008. After removing rare species following the recommendations of Piirainen et al. (2023), the final dimension of the dataset is $(N, J) = (122, 1283)$, where N is the number of species and J the number of observation sites. We apply bi-cross-validation (Owen and Perry, 2009; Owen and Wang, 2016) on the dataset and determine the number of latent features as $K = 4$.

Given the data sparsity, we threshold item parameters based on their asymptotic distribution, following Section 6.2. Figure 7 shows the full species-by-feature assignment matrix; Figure 8 displays the spatial map of the estimated item-parameter support. Next, we sum-

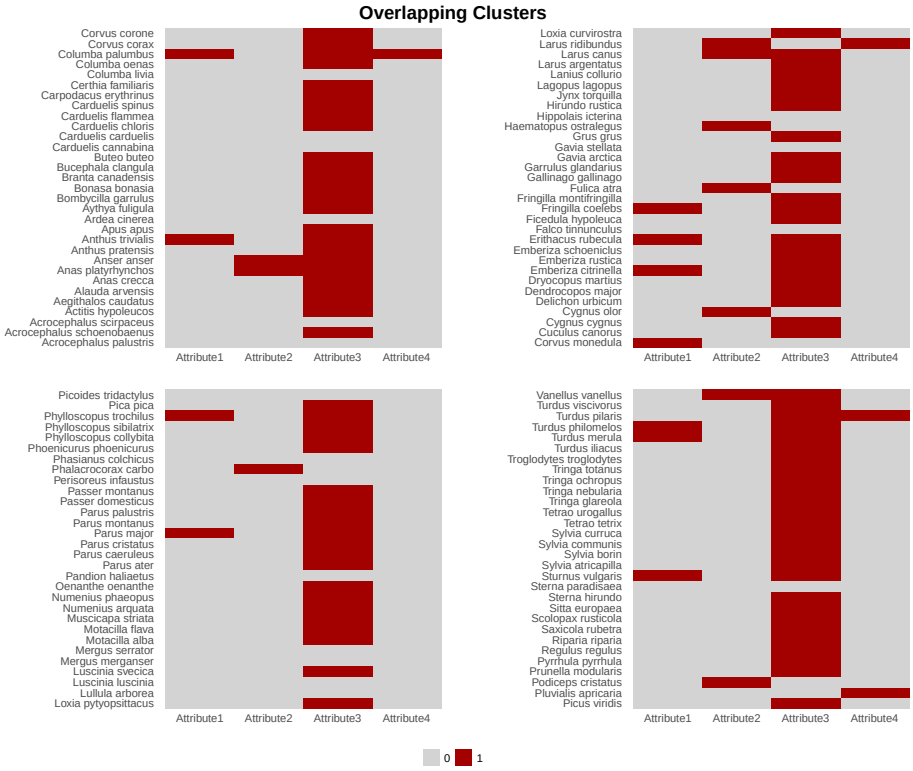

Figure 7: Estimated ecology species-by-feature assignment matrix. Red entries indicate species assigned to a recovered feature; grey entries indicate no assignment.

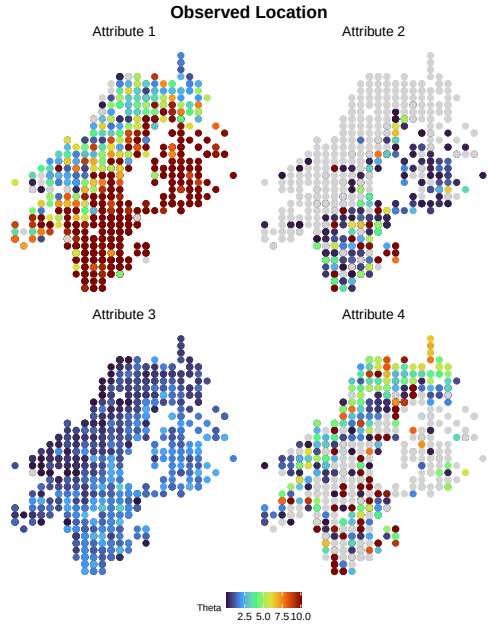

Figure 8: Spatial support of the estimated item parameters for the four recovered ecological attributes. Red points indicate locations with nonzero estimated support for the corresponding attribute; grey points indicate zero support.

marize the dominant taxonomic and spatial patterns associated with each recovered feature.

In the fitted model, the species-by-feature structure aligns with known species characteristics. To form a basis for interpretation, we compare the estimated binary latent features with species background information, focusing on phenotypic traits and migratory status. The fitted structure assigns no attributes to 18 species, 17 of which are classified as critically endangered, endangered, vulnerable, or near threatened. Regarding the specific attributes, Attribute 1 is enriched for resident birds of the order Passeriformes, whereas Attribute 2 is enriched for species from the orders Anseriformes and Charadriiformes. Attribute 3 is composed predominantly of Passeriformes. Finally, Attribute 4 is enriched for birds outside the Passeriformes order, despite Passeriformes comprising the majority of species in the data.

The estimated support of the item-parameter matrix aligns with the species interpretations above and with prior literature. Attribute 1 has support across much of Scandinavia, consistent with resident birds. Attribute 2 has support concentrated in southern coastal regions, consistent with the distributions of Anseriformes (waterfowl) and Charadriiformes (shorebirds). While the spatial pattern for Attribute 3 resembles that of Attribute 1, the associated parameter values are smaller, consistent with the overlap between resident birds and Passeriformes. Finally, Attribute 4 has support concentrated in northern Scandinavia, suggesting a borderline arctic-tundra pattern consistent with results in Zhou et al. (2025).

8 Discussion

This paper studies binary latent feature models and overlapping clustering, and develops a scalable spectral estimator with statistical guarantees for identifiability, exact recovery, parameter estimation, and inference. The central message is that the combinatorial complexity of overlapping binary latent profiles can be converted into a low-dimensional geometric problem through the population singular subspace. In particular, the population embeddings take the form of a structured parallelotope, enabling us to replace a search over exponentially many latent configurations by a geometric recovery problem in a K-dimensional spectral

domain. This insight leads to a computationally efficient method and a unified statistical theory for models both with and without intercepts.

One contribution of this work worth mentioning in the context of cognitive diagnostic modeling is the connection between the asymptotic inference theory and Q-matrix assessment. Existing Q-matrix estimation methods (Liu et al., 2012; Chen et al., 2015; Xu and Shang, 2018) typically rely on likelihood-based procedures whose computational cost scales exponentially with the number of attributes K, and formal type-I error guarantees are largely unavailable in the high-dimensional regime. By contrast, the asymptotic normality established in Theorem 5 supports entrywise and joint loading-support tests on an item-by-item basis: for any fixed item j and subset of attributes $\mathcal{I} \subseteq [K]$, one can test $H_0 : \theta_{j,\mathcal{I}} = \mathbf{0}$ at a controlled type-I error level, conditional on the exact recovery of $\mathbf{A}^*$ guaranteed by Theorems 1 and 3. These item-level tests provide a computationally scalable route to data-driven Q-matrix estimation. We note, however, that the current theory does not provide simultaneous error control over the full $J \times K$ Q-matrix; extending the inference framework to cover multiplicity-adjusted guarantees such as FWER or FDR control across all item-attribute pairs remains an important direction for future work.

Several additional future extensions are of interest. First, the current methods and theoretical results are developed under independent noise and linear mean structure. It would be valuable to extend the framework to more general dependence structures or models with nonlinear link functions. Second, the theory assumes that the latent dimension is known. Developing principled procedures for selecting the number of latent features, especially in overlapping settings, would substantially improve practical applicability. Third, while the real data analyses demonstrate that the recovered features are interpretable, the current paper does not attempt a formal study of model misspecification. Understanding the robustness of the geometric structure and the downstream inference under departures from the assumed latent feature model is another important direction.

Supplementary Material. The Supplementary Material contains the proofs of all theoretical results and additional numerical results.

References

Bhattacharya, S., Fan, J., and Hou, J. (2023). Inferences on mixing probabilities and ranking in mixed-membership models. *arXiv preprint arXiv:2308.14988*.

Chen, L., Huang, C., and Gu, Y. (2024). Generalized grade-of-membership estimation for high-dimensional locally dependent data. *arXiv preprint arXiv:2412.19796*.

Chen, Y., Chi, Y., Fan, J., Ma, C., et al. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806.

Chen, Y., Culpepper, S., and Liang, F. (2020a). A sparse latent class model for cognitive diagnosis. *Psychometrika*, 85(1):121–153.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika*, 82(3):660–692.

Chen, Y., Li, X., and Zhang, S. (2020b). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115(532):1756–1770.

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, 110(510):850–866.

Condat, L. (2016). Fast projection onto the simplex and the l 1 ball. *Mathematical Programming*, 158(1):575–585.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2):179–199.

Doshi-Velez, F. and Ghahramani, Z. (2009). Correlated non-parametric latent feature models. *Uncertainty in Artificial Intelligence*.

Frank, M., Streich, A. P., Basin, D., and Buhmann, J. M. (2012). Multi-assignment clustering for Boolean data. *Journal of Machine Learning Research*, 13:459–489.

Gillis, N. and Vavasis, S. A. (2013). Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714.

Gu, Y. and Dunson, D. B. (2023). Bayesian pyramids: Identifiable multilayer discrete latent structure models for discrete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):399–426.

Gu, Y. and Xu, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research*, 20(115):1–58.

Gu, Y. and Xu, G. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics*, 48(4):2082–2107.

Gu, Y. and Xu, G. (2023). A joint MLE approach to large-scale structured latent attribute analysis. *Journal of the American Statistical Association*, 118(541):746–760.

Heller, K. A. and Ghahramani, Z. (2007). A nonparametric Bayesian approach to modeling overlapping clusters. In *Artificial Intelligence and Statistics*, pages 187–194. PMLR.

Jin, J. (2015). Fast community detection by score. *The Annals of Statistics*, 43(1):57–89.

Jin, J., Ke, Z. T., and Luo, S. (2024). Mixed membership estimation for social networks. *Journal of Econometrics*, 239(2):105369.

Ke, Z. T. and Wang, M. (2024). Using SVD for topic modeling. *Journal of the American Statistical Association*, 119(545):434–449.

Latouche, P., Birmelé, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *Annals of Applied Statistics*, 5(1):309–336.

Liu, J., Xu, G., and Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied psychological measurement*, 36(7):548–564.

Löffler, M., Zhang, A. Y., and Zhou, H. H. (2021). Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530.

Lyu, Z., Chen, L., and Gu, Y. (2025). Degree-heterogeneous latent class analysis for high-dimensional discrete data. *Journal of the American Statistical Association*, 120(552):2435–2448.

Ma, W. and de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93(14):1–26.

Mao, X., Sarkar, P., and Chakrabarti, D. (2021). Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association*, 116(536):1928–1940.

Ni, Y., Müller, P., and Ji, Y. (2020). Bayesian double feature allocation for phenotyping with electronic health records. *Journal of the American Statistical Association*, 115(532):1620–1634.

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., et al. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological monographs*, 89(3):e01370.

OECD (2023). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. PISA. OECD Publishing, Paris.

Ovaskainen, O. and Abrego, N. (2020). *Joint species distribution modelling: With applications in R*. Cambridge University Press.

Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *The Annals of Applied Statistics*, 3(2):564 – 594.

Owen, A. B. and Wang, J. (2016). Bi-Cross-Validation for Factor Analysis. *Statistical Science*, 31(1):119 – 139.

Piirainen, S., Lehikoinen, A., Husby, M., Kålås, J. A., Lindström, Å., and Ovaskainen, O. (2023). Species distributions models may predict accurately future distributions but poorly how distributions change: A critical perspective on model validation. *Diversity and Distributions*, 29(5):654–665.

Rupp, A. A. and Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, 6(4):219–262.

Scherting, B. and Dunson, D. B. (2024). Inferring latent structure in ecological communities via barcodes. *arXiv preprint arXiv:2412.08793*.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, pages 345–354.

Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, 11(3):287.

von Davier, M. and Lee, Y.-S. (2019). Handbook of diagnostic classification models. *Cham: Springer International Publishing*.

Xu, G. and Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, 113(523):1284–1295.

Zhang, A. Y. and Zhou, H. Y. (2024). Leave-one-out singular subspace perturbation analysis for spectral clustering. *The Annals of Statistics*, 52(5):2004–2033.

Zhang, S., Liu, J., and Ying, Z. (2023). Statistical applications to cognitive diagnostic testing. *Annual Review of Statistics and Its Application*, 10(1):651–675.

Zhang, Y., Levina, E., and Zhu, J. (2020). Detecting overlapping communities in networks using spectral methods. *SIAM Journal on Mathematics of Data Science*, 2(2):265–283.

Zhou, Y., Gu, Y., and Dunson, D. B. (2025). Bayesian deep latent class regression. *arXiv preprint arXiv:2503.17531*.