Disentangling Direct and Indirect Interactions in Polytomous Item Response Theory Models

Frank Nussbaum[1,2*] and **Joachim Giesen**[1]
[1] Friedrich-Schiller University, Jena, Germany
[2] DLR Institute of Data Science, Jena, Germany
{frank.nussbaum, joachim.giesen}@uni-jena.de

Abstract

Measurement is at the core of scientific discovery. However, some quantities, such as economic behavior or intelligence, do not allow for direct measurement. They represent latent constructs that require surrogate measurements. In other scenarios, non-observed quantities can influence the variables of interest. In either case, models with latent variables are needed. Here, we investigate fused latent and graphical models that exhibit continuous latent variables and discrete observed variables. These models are characterized by a decomposition of the pairwise interaction parameter matrix into a group-sparse component of direct interactions and a low-rank component of indirect interactions due to the latent variables. We first investigate when such a decomposition is identifiable. Then, we show that fused latent and graphical models can be recovered consistently from data in the high-dimensional setting. We support our theoretical findings with experiments on synthetic and real-world data from polytomous item response theory studies.

1 Introduction

In this work, we study probabilistic models that are motivated by *item response theory* (IRT), see [Hambleton *et al.*, 1991], and its applications in the social sciences. The goal of IRT is to indirectly measure latent personality traits such as economic behavior, intelligence, or well-being by using questionnaires. While classical IRT only considers the dichotomized outcomes *right* and *wrong* for each question, *polytomous* IRT, see [Ostini and Nering, 2006], allows more general discrete outcomes. Apart from right and wrong there can, for instance, be an additional *no-choice* option. Alternatively, all available options from multiple-choice questions can be taken into account. Hence, in general, IRT considers models with observed variables x from a *discrete* sample space $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i$, where the $\mathcal{X}_i = \{0, \ldots, m_i\}$ are finite sets of choice options, and it additionally considers latent variables z from a *continuous* sample space $\mathcal{Z} = \mathbb{R}^r$. Here, the number r of latent variables is small in comparison to the number d of observed variables. Moreover, all variables are assumed to be random.

Since the assumption behind IRT is that the observed outcomes can be explained by the latent traits, any probabilistic IRT model must describe the interaction between the observed and the latent variables. The simplest (unnormalized) model that respects this requirement is given by

$$p(x, z) \propto \exp\left\{ z^\top R\,\overline{x} - \frac{1}{2} z^\top z \right\},$$

where the pairwise interaction between the observed and latent variables is modeled by the $r \times m$ matrix R. Here, $m = \sum_{i=1}^{d} m_i$ and we do not use the observed variables directly but encode them as concatenated indicator variables

$$\overline{x} = \left(\{\mathbf{1}[x_1 = k]\}_{k \in [m_1]}, \ldots, \{\mathbf{1}[x_d = k]\}_{k \in [m_d]} \right) \in \{0,1\}^m,$$

where $[n] = \{1, \ldots, n\}$ for $n \in \mathbb{N}$. In general, IRT studies strive for *independent* measurements of the latent traits, that is, the observed variables should be conditionally independent given the latent variables. However, as [Chen *et al.*, 2018] have shown for classical dichotomous IRT, this assumption is often violated, resulting in potentially unjustified biases. It is safe to assume that the same also holds true for polytomous IRT. Hence, a more reasonable probabilistic model for IRT is a pairwise *conditional Gaussian* (CG) distribution, see [Lauritzen, 1996]. It is given by

$$p(x, z) \propto \exp\left\{ \frac{1}{2} \overline{x}^\top S\,\overline{x} + z^\top R\,\overline{x} - \frac{1}{2} z^\top z \right\},$$

where $S \in \mathrm{Sym}(m)$ are the *symmetric* direct interactions of the observed variables. From this joint model we can derive the marginal model for the observed discrete variables by integrating out the latent variables z as

$$\begin{aligned} p(x) &\propto \exp\left\{ \frac{1}{2} \overline{x}^\top (S + R^\top R)\,\overline{x} \right\} \\ &=: \exp\left\{ \frac{1}{2} \overline{x}^\top (S + L)\,\overline{x} \right\}, \end{aligned} \qquad (1)$$

where $L = R^\top R$ has rank at most r. This shows that a small number of latent variables induces a pairwise *low-rank* interaction between the observed variables. We still assume that

[*] Contact Author

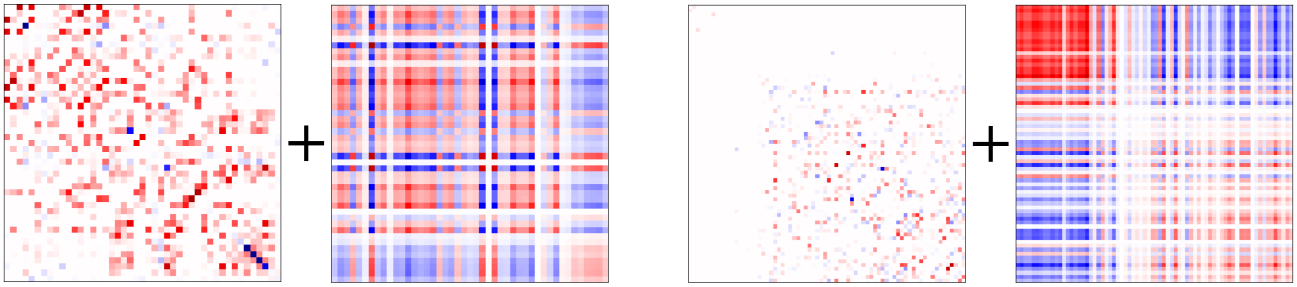

Figure 1: Learned group-sparse + low-rank decompositions for the VIQT (left) and the CFMT datasets (right) that are introduced in Section 5. The group-sparse components correspond to direct local dependencies of the observed discrete variables, and the low-rank components represent indirect effects due to the latent continuous variables. Here, red indicates positive and blue indicates negative (conditional) correlations.

the conditional independence assumption of IRT is not violated much, that is, the observed variables are mostly independent conditioned on the latent variables. Here, any conditional dependence between two observed discrete variables is determined by a *group* of parameters within the matrix S of direct interactions. These groups are given as

$$S = \left(\begin{array}{c|c|c|c} S_{11} & S_{12} & \cdots & S_{1d} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline S_{d1} & S_{d2} & \cdots & S_{dd} \end{array} \right) \in \mathrm{Sym}(m), \qquad (2)$$

where the group $S_{ij} \in \mathbb{R}^{m_i \times m_j}$ contains the interaction parameters for the i-th and j-th observed variable. Hence, the absence of a direct pairwise interaction between two observed variables means that *all* parameters in the corresponding group are zero. Consequently, the assumption that most of the observed variables are conditionally independent implies *group sparsity* of S. Overall, the interaction between the observed variables in our marginal probabilistic IRT model is the sum of a group-sparse and a low-rank matrix, see Figure 1 for two examples.

For the proper analysis of IRT models, it is important to disentangle the group-sparse and low-rank interactions because only the low-rank interactions are induced by the latent variables, which are of primary interest here. Hence, the goal of this work is the design and analysis of a method for *consistent* estimation of the group-sparse and the low-rank part of the interaction matrix for the observed variables. Here, we use the classical maximum likelihood approach for estimating the interaction parameter matrix $\Theta = S + L$ of the discrete variables. To simplify our exposition, we *minimize* the convex *negative* log-likelihood function that can be derived from our model and reads as

$$\ell(\Theta) = a(\Theta) - \mathrm{tr}(\Theta\Sigma^n),$$

where $a(\Theta)$ is the *log-partition* (normalization) function of the model and $\Sigma^n = 1/n \sum_{k=1}^{n} \overline{x}^{(k)} \otimes \overline{x}^{(k)}$ is the *empirical second-moment matrix* that is computed as the average over outer products of n indicator-encoded observations $\overline{x}^{(k)} \in \{0,1\}^m$. Since by our assumptions the interaction matrix Θ is the sum of a group-sparse and a low-rank matrix, we promote this structure in the estimation process by adding group-sparsity and low-rank inducing regularization terms to the objective. First, it is well known that low rank can be induced on positive semidefinite L by trace-norm regularization $\mathrm{tr}(L)$. Second, group-sparse structure can be induced on S by $\ell_{1,2}$-norm regularization. Here, the $\ell_{1,2}$-norm is given by $\|S\|_{1,2} = \sum_{i,j=1}^{d} \|S_{ij}\|_2$, where, depending on the dimensions of S_{ij}, the norm $\|S_{ij}\|_2$ denotes the absolute value, the Euclidean vector norm, or the Frobenius norm. Hence, we study the following convex optimization problem

$$\min_{S,\, L \succeq 0} \ \ell(S + L) + \lambda \left(\gamma \|S\|_{1,2} + \mathrm{tr}(L)\right) \qquad (3)$$

that uses regularization parameters $\lambda, \gamma > 0$. In Section 4, we will show that under some mild technical conditions, the solution to Problem (3) with appropriate regularization parameters can indeed consistently disentangle the group-sparse and the low-rank components of the interaction matrix.

2 Contributions and Related Work

A necessary condition for successful disentanglement of the group-sparse and low-rank components of the interaction matrix is that our model is *identifiable*. A parameterized class of probability distributions is identifiable if no distribution from the class has two different parameterizations. In our case this means that there should be no two group-sparse matrices S and S' and two low-rank matrices L and L' with $(S, L) \neq (S', L')$ such that $S + L = \Theta = S' + L'$. In Sections 3 and 4, we work out conditions that ensure identifiability of our model.

The issue of identifiability received considerable attention in the context of *mixture models*, where the sample space $\mathcal{Z}$ of the latent variables is, in contrast to our model, a finite set. In this setting, the marginal distribution on the sample space $\mathcal{X}$ of the observed variables is a mixture of distributions $p(x) = \sum_{z \in \mathcal{Z}} p(x, z) = \sum_{z \in \mathcal{Z}} p(z)p(x|z)$, where $p(x|z)$ are the mixture components with corresponding weights $p(z)$. A mixture model can be interpreted as a clustering of the observed samples into $|\mathcal{Z}|$ clusters, where the posterior $p(z|x)$ is the probability that x belongs to cluster $z \in \mathcal{Z}$. One must however be cautious with such an interpretation if there exists a different *global* model that yields the same marginal model. In this case, there is an identifiability problem since the joint models cannot be told apart from what can be observed, although they describe different clusterings. [Carreira-Perpinán and Renals, 2000], [Allman *et al.*, 2009], and [Montúfar and Morton, 2017] provide conditions that ensure the identifiability of various mixture models.

Closer to our setting, [Candès *et al.*, 2011] and [Chandrasekaran *et al.*, 2011] provide conditions for the identifiability of matrix decompositions into sparse and low-rank components. [Tang and Nehorai, 2011] extend these conditions to column-sparse + low-rank matrix decompositions. In this work, we generalize the previous identifiability results to the group-sparse case that we motivated before.

Next, recall that a learning method is called *consistent* if it can recover the parameters of a probabilistic model in the limit of a growing number of data points that have been sampled from the model. Note that there cannot be a consistent learning method for non-identifiable models and thus the existence of a consistent learning method is a stronger property. [Ravikumar *et al.*, 2011] and [Jalali *et al.*, 2011], respectively, show that the parameters of Gaussian and discrete graphical models can be recovered consistently via convex optimization. In pioneering work, [Chandrasekaran *et al.*, 2012] extend this approach to latent variable Gaussian graphical models, that is, to multivariate Gaussians with observed and latent variables. In previous work it was shown that consistent recovery is also possible for sparse + low-rank models with observed binary variables [Nussbaum and Giesen, 2019] and with mixed binary and conditional Gaussian observed variables [Nussbaum and Giesen, 2020]. Here, we extend their approach to group-sparse + low-rank interaction matrices. In Section 5, we corroborate our theoretical findings in experiments on synthetic data. Furthermore, we demonstrate the expediency of our model and model-selection approach on two polytomous IRT datasets from the social sciences.

3 Identifiability of the Model

In this section, we discuss the identifiability of the decomposition of the pairwise interaction parameter matrix from Model (1). For that, we consider group-sparse matrices that are contained in the *group-sparse matrix variety* of symmetric matrices with at most s non-zero groups given by

$$\mathcal{S}(s) = \{S \in \operatorname{Sym}(m) : |\operatorname{gsupp}(S)| \leq s\},$$

where $\operatorname{gsupp}(S) = \{(i,j) \in [d] \times [d] :\ S_{ij} \not\equiv 0\}$ defines the *group support* of S. Here, the groups $S_{ij} \in \mathbb{R}^{m_i \times m_j}$ are as in Equation (2). Moreover, we consider low-rank matrices from the *low-rank matrix variety* of symmetric matrices with rank at most r that is given by

$$\mathcal{L}(r) = \{L \in \operatorname{Sym}(m) :\ \operatorname{rank}(L) \leq r\}.$$

Next, we provide a condition that ensures *local* identifiability of the product variety $\mathcal{S}(s) \times \mathcal{L}(r)$ for fixed s and r. We call $(S, L) \in \mathcal{S}(s) \times \mathcal{L}(r)$ locally identifiable if $(S+\Delta, L-\Delta) \notin \mathcal{S}(s) \times \mathcal{L}(r)$ for all $\Delta \neq 0$ from some small ball. Hence, we need to characterize nearby points in the varieties. First, if $S \in \mathcal{S}(s)$ with $|\operatorname{gsupp}(S)| = s$, then $S + \Delta \in \mathcal{S}(s)$ for small Δ if and only if Δ is contained in the tangent space

$$\mathcal{Q}(S) = \{M \in \operatorname{Sym}(m) :\ \operatorname{gsupp}(M) \subseteq \operatorname{gsupp}(S)\}$$

at S to $\mathcal{S}(s)$. Second, a rank-r matrix L has the tangent space

$$\mathcal{T}(L) = \left\{UX^\top + XU^\top :\ X \in \mathbb{R}^{m \times r}\right\} \subset \operatorname{Sym}(m)$$

to $\mathcal{L}(r)$, where $L = UDU^\top$ is the (restricted) eigenvalue decomposition of L with $U \in \mathbb{R}^{m \times r}$ and $D \in \mathbb{R}^{r \times r}$. This time, because the low-rank matrix variety is locally curved, having $L - \Delta \in \mathcal{L}(r)$ for small Δ only implies that Δ is a direction from some tangent space $\mathcal{T}(L')$ to $\mathcal{L}(r)$ at a matrix $L' \in \mathcal{L}(r)$ that is close to L. The following lemma shows that it is still sufficient to only consider $\mathcal{T}(L)$ for local identifiability.

Lemma 1. *Let $\mathcal{Q}(S) \cap \mathcal{T}(L) = \{0\}$, that is, we assume that the tangent spaces $\mathcal{Q}(S)$ and $\mathcal{T}(L)$ are* transverse. *Then, (S, L) is locally identifiable in $\mathcal{S}(s) \times \mathcal{L}(r)$, where $s = |\operatorname{gsupp}(S)|$ and $r = \operatorname{rank}(L)$.*

To establish this result, one can prove that transversality of the tangent spaces extends to nearby tangent spaces. Now, restricting $\mathcal{S}(s) \times \mathcal{L}(r)$ to points (S, L) with $\mathcal{Q}(S) \cap \mathcal{T}(L) = \{0\}$ leads to a class of locally identifiable models, though not globally identifiable. For example, using the first and m-th standard basis vectors of $\mathbb{R}^m$, the pair consisting of $S = e_1 e_1^\top$ and $L = e_m e_m^\top$ is locally identifiable in $\mathcal{S}(1) \times \mathcal{L}(1)$ since for $d \geq 2$ it holds $\mathcal{Q}(S) \cap \mathcal{T}(L) = \{0\}$. However, exchanging the roles of S and L yields a different parametrization $(L, S) \in \mathcal{S}(1) \times \mathcal{L}(1)$ of the same model.

The problem in the example is that *both* components are group sparse and low rank and thus can be confused. To avoid this, first the matrix S should have a small maximum degree $\operatorname{gdeg}_{\max}(S)$, that is, a small maximum number of non-zero groups per row/column. If this is the case, S cannot be mistaken as low rank. Second, the low-rank matrix L should have a row/column space $U(L) \subseteq \mathbb{R}^m$ that is not well-aligned with the standard-basis vectors e_i of $\mathbb{R}^m$. Formally, the *incoherence* $\operatorname{coh}(L) = \max_i \|P_{U(L)} e_i\|_2$ of L should be small because then L is spread-out and cannot be confused with a group-sparse matrix. The next lemma shows that bounding the product $\operatorname{gdeg}_{\max}(S) \operatorname{coh}(L)$ implies tangent space transversality and thus identifiability by Lemma 1.

Lemma 2. *Let $(S, L) \in \mathcal{S}(|\operatorname{gsupp}(S)|) \times \mathcal{L}(\operatorname{rank}(L))$ and $\eta = \max_{i \in [d]} m_i$. Then, provided that it holds that $\operatorname{gdeg}_{\max}(S) \operatorname{coh}(L) < 1/2\eta^{-3/2}$, the tangent spaces $\mathcal{Q}(S)$ and $\mathcal{T}(L)$ are transverse, that is, $\mathcal{Q}(S) \cap \mathcal{T}(L) = \{0\}$.*

Observe that for (S, L) as in the previously discussed example it holds $\operatorname{gdeg}_{\max}(S) \operatorname{coh}(L) = 1$ such that the condition of Lemma 2 is not satisfied. In the next section, we will see that a slightly stronger upper bound on the product $\operatorname{gdeg}_{\max}(S) \operatorname{coh}(L)$ even allows the consistent recovery of (S, L) by solving instances of Problem (3).

4 Consistency Analysis

For our consistency analysis we assume that n data points have been drawn from Model (1) with *true* interaction parameter matrix $S^\star + L^\star$, where $S^\star$ is group sparse and $L^\star$ is low rank. We show that the solution (S_n, L_n) to Problem (3) can recover the true parameters $(S^\star, L^\star)$ consistently, that is, asymptotically and with high probability. Specifically, we show two types of consistency. The first type is *algebraic consistency*. It holds if S_n has the same group support as $S^\star$ and if L_n has the same rank as $L^\star$. The second type is *parametric consistency*, which holds if the errors $S_n - S^\star$ and $L_n - L^\star$ are small. Following [Chandrasekaran *et al.*, 2012], a good measure for the size of the errors is the dual norm of the regularizing norm $\gamma\|S\|_{1,2} + \|L\|_*$ from the objective function. We call this dual norm the γ-*norm*. It is given by

$$\|(M,N)\|_\gamma = \max\left\{\gamma^{-1}\|M\|_{\infty,2}, \|N\|\right\}$$

for $(M,N) \in \mathrm{Sym}(m) \times \mathrm{Sym}(m)$, where $\|M\|_{\infty,2} = \max_{i,j}\|M_{ij}\|_2$, and $\|N\|$ is the spectral norm of N.

Consistent recovery is not always possible. A first challenge is controlling the sampling error, which is given by $\nabla\ell(S^\star + L^\star) = \nabla a(S^\star + L^\star) - \Sigma^n = \mathbb{E}[\Sigma] - \Sigma^n$, where the expectation is w.r.t. the true model. If the sampling error is small, then the likelihood term in Problem (3) ensures that the compound matrix $S_n + L_n$ is close to the true compound matrix $S^\star + L^\star$. However, reliable recovery of the compound matrix is only possible if λ is not too large. Indeed, the regularization terms should only encourage small adjustments to the algebraic structure of the components. Hence, later we assume an upper bound on λ.

A second challenge is telling the group-sparse and low-rank components apart. This can be addressed by restricting the analysis to identifiable models as outlined in Section 3. For a better understanding, let us consider the intuitive variety-constrained version

$$\min \quad \ell(S+L) \quad \text{s.t.} \quad S \in \mathcal{S}(s),\, L \in \mathcal{L}(r)$$

of Problem (3). This non-convex problem is of a hypothetical nature because the true group-sparse and low-rank matrix varieties with $s = |\operatorname{gsupp}(S^\star)|$ and $r = \operatorname{rank}(L^\star)$ are unknown when solving the problem. Nevertheless, Problem (3) can be seen as a convex relaxation of the variety-constrained problem. Therefore intuitively, the solution of the variety-constrained problem should be (locally) unique in order to successfully recover the true parameters. The main reason for local non-uniqueness can be a decomposition $S + L$ that is not locally identifiable. This can be excluded by Lemma 1 provided that $\mathcal{Q}(S) \cap \mathcal{T}(L) = \{0\}$. However, we must also ensure that no nearby solutions with different compound matrices exists. To fully characterize local uniqueness, we use the optimality conditions of the variety-constrained problem, see Figure 2. They state that (S, L) can only be a (local) solution if the gradient of the negative log-likelihood at $S + L$ is *normal* to the respective varieties, that is, if it holds that

$$\nabla\ell(S+L) \perp \mathcal{Q}(S) \quad \text{and} \quad \nabla\ell(S+L) \perp \mathcal{T}(L).$$

Here, the gradient is the same with respect to S and L. To guarantee local uniqueness, the optimality condition should be violated at any *slightly perturbed* solution, that is, the gradient at such a perturbed solution should be non-normal to at least one of the varieties.

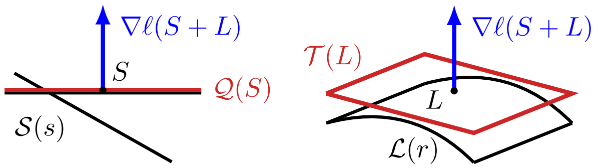

Figure 2: Optimality conditions for a solution (S, L) to the variety-constrained problem: The gradient must be normal to the group-sparse matrix variety $\mathcal{S}(s)$ and the low-rank matrix variety $\mathcal{L}(r)$.

In what follows, we focus on perturbations in tangential directions since the normal spaces to the varieties hardly change for such perturbations. Hence, for a tangentially perturbed solution, the optimality conditions are surely violated *if* the gradient at the perturbed solution has components in the tangent spaces that are large compared to the components in the normal spaces. Therefore, we need to control the change of the gradient under small (tangential) perturbations Δ. It is sufficient to do this only for perturbations from the true solution $(S^\star, L^\star)$ since our assumptions will carry over to any nearby local solution (S, L). Now, given a small sampling error $\nabla\ell(S^\star + L^\star)$, the change of the gradient is locally governed by the Hessian $H^\star = \nabla^2\ell(S^\star + L^\star)$. This follows from the first-order approximation

$$\begin{aligned}\nabla\ell(S^\star + L^\star + \Delta) &\approx \nabla\ell(S^\star + L^\star) + \nabla^2\ell(S^\star + L^\star)\Delta \\ &\approx \nabla^2\ell(S^\star + L^\star)\Delta = H^\star\Delta.\end{aligned}$$

Consequently, the tangential components of the gradient at the perturbed solution are large if the *minimum gains* of the Hessian $H^\star$ in tangential directions are large. At the same time, the normal components of the gradient are not too large if the *orthogonal effects* of the Hessian $H^\star$ in normal directions are not too large compared to the tangential effects. We formalize this in our main assumption.

Assumption 1. *We assume that there exist $\alpha > 0$, $\nu \in (0, 1/2]$, and an interval $[\gamma_{\min}, \gamma_{\max}]$ such that the following holds for any $\gamma \in [\gamma_{\min}, \gamma_{\max}]$, any (low-rank) tangent space $\mathcal{T}(L')$ that is sufficiently close to $\mathcal{T}(L^\star)$, and any $(M, N) \in \mathcal{J} = \mathcal{Q}(S^\star) \times \mathcal{T}(L')$: First, the minimum gain on $\mathcal{J}$ of $H^\star$ is bounded from below in the sense that*

$$\|P_{\mathcal{J}} DH^\star(M+N)\|_\gamma \ge \alpha/2\,\|(M,N)\|_\gamma\,,$$

where $D : \mathrm{Sym}(m) \to \mathrm{Sym}(m) \times \mathrm{Sym}(m), M \mapsto (M, M)$ is the duplication operator. Second, the orthogonal effect on $\mathcal{J}^\perp$ of $H^\star$ is bounded from above in the sense that

$$\|P_{\mathcal{J}^\perp} DH^\star(M+N)\|_\gamma \le (1-\nu)\,\|P_{\mathcal{J}} DH^\star(M+N)\|_\gamma\,.$$

A few comments are in order to understand this assumption. First, note that Assumption 1 is a generalization of the irrepresentability assumption as for example used by [Zhao and Yu, 2006, Wainwright, 2009, Ravikumar *et al.*, 2011]. Second, Assumption 1 concerns tangent spaces that are close to $\mathcal{T}(L^\star)$ to account for the local curvature of the low-rank matrix variety. Third, the assumption on the minimum gain implies transversality, that is, $\mathcal{Q}(S^\star) \cap \mathcal{T}(L') = \{0\}$. In conjunction with Lemma 1, this entails local identifiability. Fourth and finally, it can be shown that the existence of the interval $[\gamma_{\min}, \gamma_{\max}]$ is implied by the upper bound $\operatorname{gdeg}_{\max}(S^\star)\operatorname{coh}(L^\star) \le c/12\,\eta^{-3/2}(\alpha\nu)^2(2-\nu)^{-2}$ that uses the constants α and ν from the assumption as well as another constant c that is independent of n and d. Recall that in Section 3 we have seen that *small* $\operatorname{gdeg}_{\max}(S^\star)$ and $\operatorname{coh}(L^\star)$ help to avoid confusion of $S^\star$ and $L^\star$. Hence, assuming an upper bound on $\operatorname{gdeg}_{\max}(S^\star)\operatorname{coh}(L^\star)$ is reasonable.

Next, we formulate our consistency result, assuming a small sampling error first. We control the sampling error by a probabilistic analysis afterwards. Moreover, for better readability, we leave out constants that are independent of n, d, the group sparsity of $S^\star$, and the coherence of $L^\star$. We denote inequalities that hold up to such constants by $\lesssim$ and $\gtrsim$.

Theorem 1. *Let $S^\star, L^\star \in \mathrm{Sym}(m)$ with $L^\star \succeq 0$ such that Assumption 1 is satisfied. Suppose that we observed samples $x^{(1)}, \dots, x^{(n)} \in \mathcal{X}$ drawn from Model (1) with interaction matrix $S^\star + L^\star$. Assume that $\lambda \lesssim \mathrm{coh}(L^\star)$, that $\gamma \in [\gamma_{\min}, \gamma_{\max}]$, and that the sampling error is bounded as*

$$\|D\nabla\ell(S^\star + L^\star)\|_\gamma = \|D(\mathbb{E}[\Sigma] - \Sigma^n)\|_\gamma \lesssim \lambda.$$

Also assume that the minimum magnitude of the non-zero groups of $S^\star$ is bounded from below as $s_{\min} = \min_{(i,j)\,\in\,\mathrm{gsupp}(S^\star)} \|S_{ij}\|_2 \gtrsim \lambda |\,\mathrm{gsupp}(S^\star)|^{-1} d$ and that the minimum non-zero singular value of $L^\star$ is bounded from below as $\sigma_{\min} \gtrsim \lambda\, \mathrm{coh}(L^\star)^{-1}$. Then, it follows that the solution (S_n, L_n) to Problem (3) with regularization parameters λ and γ is unique and

(a) parametrically consistent in the sense that it satisfies $\|(S_n - S^\star, L_n - L^\star)\|_\gamma \lesssim \lambda$ and

(b) algebraically consistent, that is, S_n and $S^\star$ have the same group support, and L_n and $L^\star$ have the same rank.

We make a few additional comments concerning the assumptions of Theorem 1. First, recall that an upper bound on λ is needed for reliable recovery of the compound matrix $S^\star + L^\star$. We also assume lower bounds on $s_{\min}$ and $\sigma_{\min}$ in terms of λ to avoid that the shrinkage effects due to the regularization cause algebraic features of the decomposition to disappear. Next, it is natural to require the sampling error to be bounded by λ since intuitively, for large λ and thus strong regularization, we can allow for a larger sampling error. However, while the probability that the sampling-error bound holds for an actual sample will also be larger for large λ, the error bound in Theorem 1(a) gets weaker.

Similarly as in [Chandrasekaran *et al.*, 2012] and [Nussbaum and Giesen, 2020], Theorem 1 can be proven using the *primal-dual witness* technique: It proceeds by first restricting Problem (3) to a (non-convex) correct model set $\mathcal{M}$ that is chosen in a way such that any solution $(S_\mathcal{M}, L_\mathcal{M})$ to the restricted problem is algebraically and parametrically consistent. The non-convexity is due to a rank constraint, which is subsequently replaced by a linear tangent-space constraint to the low-rank matrix variety at a fixed solution $(S_\mathcal{M}, L_\mathcal{M})$. Then, it is shown that the solution to the linearized problem is unique and coincides with $(S_\mathcal{M}, L_\mathcal{M})$. Finally, it is shown that the original Problem (3) is also solved by the same consistent solution $(S_\mathcal{M}, L_\mathcal{M})$.

Next, we show that the bound on the sampling error that is required by Theorem 1 holds with high probability for a specific choice of λ.

Corollary 1. *Under the assumptions of Theorem 1 let*

$$\lambda = \lambda_{n,d} \asymp \frac{\eta}{\mathrm{coh}(L^\star)} \sqrt{\frac{\kappa d \log m}{n}}$$

for some $\kappa \geq 1$ and let $n \gtrsim \kappa\eta^2\, \mathrm{coh}(L^\star)^{-4} d \log m$. Then, it follows that $\|D(\mathbb{E}[\Sigma] - \Sigma^n)\|_\gamma \lesssim \lambda_{n,d}$ with probability at least $1 - m^{-\kappa}$. Hence, the conclusions of Theorem 1 applied with $\lambda = \lambda_{n,d}$ hold with the same probability.

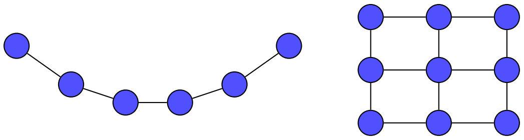

Figure 3: Conditional graph structures: chain (left) and grid (right). Note that each edge corresponds to a group of several parameters.

Corollary 1 can be proven using concentration results for *bounded* random vectors [Vershynin, 2010]. The dependence on $\mathrm{coh}(L^\star)$ can be improved in both Theorem 1 and Corollary 1 in a rather technical and non-intuitive way. However, here we choose to remain clear and self-contained. We investigate the influence of $\mathrm{coh}(L^\star)$ and $\mathrm{gdeg}_{\max}(S^\star)$ in our experiments in the next section.

The specific choice $\lambda = \lambda_{n,d}$ in Corollary 1 allows the discussion of some *high-dimensional* limits. For that we consider problems that vary in the number of data points n, in the number of observed variables d, and in the group sparsity and the incoherence of the components. Apart from that, we only consider problems that satisfy the assumptions from Theorem 1 for the same constants, particularly the same α and ν. In this setting, first observe that for fixed d the error bound $\|(S_n - S^\star, L_n - L^\star)\|_\gamma \lesssim \lambda_{n,d}$ from Corollary 1 asymptotically approaches zero as $n \to \infty$. Second, for larger d, more samples are required to obtain the same error bound. Furthermore, we expect that larger $\mathrm{gdeg}_{\max}(S^\star)$ and larger $\mathrm{coh}(L^\star)$ make consistent recovery more difficult since then it is more likely that the upper bound on $\mathrm{gdeg}_{\max}(S^\star)\,\mathrm{coh}(L^\star)$ is not satisfied. We also investigate this in our experiments.

5 Experiments

We solve Problem (3) using the Alternating Direction Method of Multipliers (ADMM) with a proximal gradient step that we adapted to our needs from [Ma *et al.*, 2013]. For computational efficiency, we replace the likelihood by the pseudo-likelihood in our experiments. It is known to behave similarly [Mozeika *et al.*, 2014].

5.1 Synthetic Data

Here, to verify experimentally that consistent recovery is possible, we generate synthetic data from discrete fused latent and graphical models using Gibbs sampling, see [Casella and George, 1992]. For the experiments, we use discrete variables that take three values, that is, $\mathcal{X}_i = \{0, 1, 2\}$ for all variables. We consider four models with $d = 36$ variables, where the direct interactions $S^\star$ adhere to either chain or grid graphical model structures (compare Figure 3), and the number of latent variables is either one or two.

Our goal is to test the influence of the maximum group degree $\mathrm{gdeg}_{\max}(S^\star)$ of $S^\star$ and the incoherence $\mathrm{coh}(L^\star)$ of $L^\star$ on recovery rates. Here, the maximum group degree is two for the chain and four for the grid model. Moreover, we set the probability of an interaction between any latent and any observed variable to be non-zero to 95%. This ensures that the low-rank effect of the latent variables is spread-out and

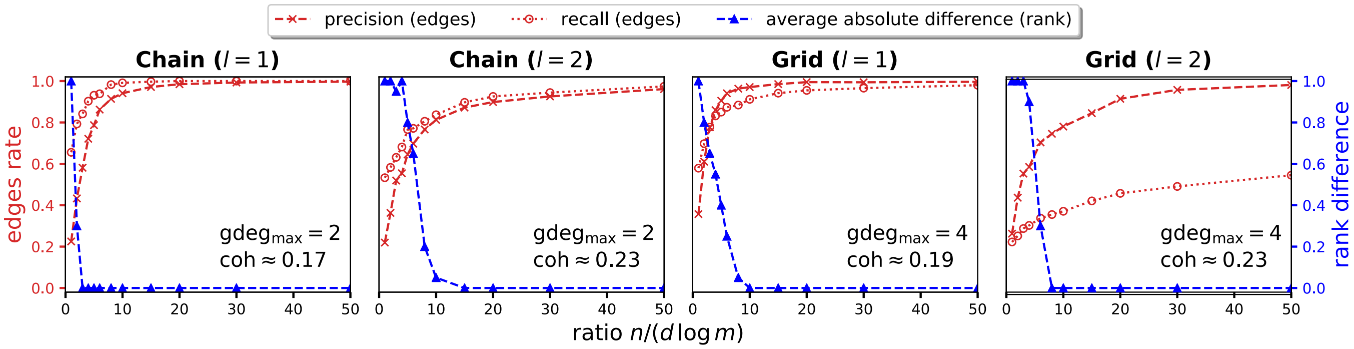

Figure 4: Recall, precision, and absolute rank difference averaged over 20 trials for each model and ratio. In the plots for each model, the maximum group degree of $S^\star$ and the coherence of the randomly sampled $L^\star$ are noted.

thus that $L^\star$ is incoherent. However, $L^\star$ will be less incoherent for a growing number of latent variables.

For each model, we sample all of its parameters randomly. More specifically, we sample the latent-observed interaction parameters uniformly from $[-0.5, -0.2] \cup [0.2, 0.5]$ and the parameters for the non-zero groups of $S^\star$ from $[-1.5, -0.5] \cup [0.5, 1.5]$. Then, for each model we test the asymptotic behavior by generating 20 datasets with $kd \log m$ samples (rounded to the nearest integer) for selected ratios $k \in [1, 50]$. Our choice of regularization parameters is guided by Corollary 1 and fixed for all models ($\lambda = 1/50\,\sqrt{d \log m/n}$, $\gamma = 10$). Finally, for each model and ratio k, we record the average percentage of correctly identified non-zero groups, that is, edges in the conditional independence graph. For that, we employ the criteria of *recall* and *precision*, where $\mathrm{recall} = \mathrm{TP}\,/(\mathrm{TP} + \mathrm{FN})$ and $\mathrm{precision} = \mathrm{TP}\,/(\mathrm{TP} + \mathrm{FP})$. Here, TP is the number of correctly identified edges (true positives), FN is the number of undetected edges (false negatives), and FP is the number of edges that were mistakenly detected as edges (false positives). Likewise, we record the absolute rank difference $|\operatorname{rank}(L_n) - \operatorname{rank}(L^\star)|$, averaged over the 20 trials.

The results are shown in Figure 4. Recovery of edges and rank requires relatively few samples for the one-latent-variable chain and grid models. Slightly more samples are required to recover the rank of the grid model. This is due to the larger maximum group degree of the grid models compared to the chain models. Next, for the two-latent-variable chain model considerably more samples are necessary for successful recovery—because the underlying low-rank matrix is less incoherent. Our observations back the intuition that for more incoherent $L^\star$ and smaller maximum group degree of $S^\star$, the group-sparse and low-rank components can be confused less easily. This is supported even more by the recovery results for the two-latent-variable grid model, where the fact that both the maximum group degree and the coherence are larger is reflected in significantly worse recovery results. Nevertheless, overall the results show that consistent recovery is possible.

5.2 Real-World Data

We also demonstrate the effectiveness of our fused latent and graphical models on two real-world datasets. The first dataset is from a *non-forced* choice vocabulary IQ test (VIQT), where participants can indicate if they do not know an answer, otherwise answers are either correct or wrong. The dataset was obtained from the [Open-Source Psychometrics Project, 2019] and contains $d = 45$ variables and $n = 12\,173$ samples. The second dataset contains the answers of $n = 165$ test takers to the $d = 72$ questions of the Cambridge face memory test (CFMT) [Duchaine and Nakayama, 2006,Itz *et al.*, 2017]. In this dataset, answers with response times below the human reaction time or above some threshold (based on the interquartile range) were assigned to a third category of outliers, otherwise answers are either correct or wrong, again. Hence, for both datasets, the observed variables are discrete with three possible outcomes.

Figure 1 shows estimated fused latent and graphical models for both datasets. The learned models exhibit direct interactions, that is, the answers are not independent given the estimated latent variables. This is in contrast to the common conditional independence assumption in item response theory. Nevertheless, for both models, most observed interactions are explained by a latent variable. Notably, the low-rank matrix learned for the CFMT data has a block of positively correlated items in the top left corner. These items correspond to the first CFMT block, which consists of 18 simple questions that most participants get right, hence the correlation.

6 Conclusion

In this paper, we studied discrete fused latent and graphical models. We investigated when the group-sparse + low-rank decomposition of the interaction parameter matrix is identifiable. We showed that using convex optimization such decompositions can be estimated consistently in the high-dimensional setting. We verified experimentally that consistent recovery becomes easier if there are not too many non-zero groups per row/column of the matrix of direct interactions and if the low-rank matrix is spread-out. The fused latent and graphical models that we estimated from real-world data demonstrate that observed data from polytomous IRT studies can be conditionally dependent given the latent variables—in contrast to the common assumption. This shows that modeling direct interactions in polytomous IRT is warranted and reasonable.

References

[Allman *et al.*, 2009] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

[Candès *et al.*, 2011] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[Carreira-Perpinán and Renals, 2000] Miguel A. Carreira-Perpinán and Steve Renals. Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Computation*, 12(1):141–152, 2000.

[Casella and George, 1992] George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

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

[Chandrasekaran *et al.*, 2012] Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

[Chen *et al.*, 2018] Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. Robust measurement via a fused latent and graphical item response theory model. *Psychometrika*, pages 1–25, 2018.

[Duchaine and Nakayama, 2006] Brad Duchaine and Ken Nakayama. The cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, 44(4):576–585, 2006.

[Hambleton *et al.*, 1991] Ronald K. Hambleton, Hariharan Swaminathan, and H. Jane Rogers. *Fundamentals of item response theory*. Sage, 1991.

[Itz *et al.*, 2017] Marlena L. Itz, Jessika Golle, Stefanie Luttmann, Stefan R. Schweinberger, and Jürgen M. Kaufmann. Dominance of texture over shape in facial identity processing is modulated by individual abilities. *British Journal of Psychology*, 108(2):369–396, 2017.

[Jalali *et al.*, 2011] Ali Jalali, Pradeep Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 378–387, 2011.

[Lauritzen, 1996] Steffen L. Lauritzen. *Graphical models*. Oxford University Press, 1996.

[Ma *et al.*, 2013] Shiqian Ma, Lingzhou Xue, and Hui Zou. Alternating direction methods for latent variable Gaussian graphical model selection. *Neural Computation*, 25(8):2172–2198, 2013.

[Meinshausen and Bühlmann, 2006] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[Montúfar and Morton, 2017] Guido Montúfar and Jason Morton. Dimension of marginals of Kronecker product models. *SIAM Journal on Applied Algebra and Geometry*, 1(1):126–151, 2017.

[Mozeika *et al.*, 2014] Alexander Mozeika, Onur Dikmen, and Joonas Piili. Consistent inference of a general model using the pseudolikelihood method. *Physical Review E*, 90(1), 2014.

[Nussbaum and Giesen, 2019] Frank Nussbaum and Joachim Giesen. Ising models with latent conditional Gaussian variables. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, 2019.

[Nussbaum and Giesen, 2020] Frank Nussbaum and Joachim Giesen. Pairwise sparse + low-rank models for variables of mixed type. *Journal of Multivariate Analysis*, 178:104601, 2020.

[Open-Source Psychometrics Project, 2019] Open psychology data: Raw data from online personality tests. https://openpsychometrics.org/_rawdata, 2019.

[Ostini and Nering, 2006] Remo Ostini and Michael L. Nering. *Polytomous item response theory models*. Number 144 in Quantitative Applications in the Social Sciences. Sage, 2006.

[Ravikumar *et al.*, 2010] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[Ravikumar *et al.*, 2011] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[Tang and Nehorai, 2011] Gongguo Tang and Arye Nehorai. Robust principal component analysis based on low-rank and block-sparse matrix decomposition. In *2011 45th Annual Conference on Information Sciences and Systems*, pages 1–5. IEEE, 2011.

[Vershynin, 2010] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. Technical report, arXiv preprint arXiv:1011.3027, 2010.

[Wainwright, 2009] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55(5):2183–2202, 2009.

[Zhao and Yu, 2006] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.