Ising Models with Latent Conditional Gaussian Variables

Frank Nussbaum FRANK.NUSSBAUM@UNI-JENA.DE
Institut für Informatik
Friedrich-Schiller-Universität Jena
Germany

Joachim Giesen JOACHIM.GIESEN@UNI-JENA.DE
Institut für Informatik
Friedrich-Schiller-Universität Jena
Germany

Editors: Aurélien Garivier and Satyen Kale

Abstract

Ising models describe the joint probability distribution of a vector of binary feature variables. Typically, not all the variables interact with each other and one is interested in learning the presumably sparse network structure of the interacting variables. However, in the presence of latent variables, the conventional method of learning a sparse model might fail. This is because the latent variables induce indirect interactions of the observed variables. In the case of only a few latent conditional Gaussian variables these spurious interactions contribute an additional low-rank component to the interaction parameters of the observed Ising model. Therefore, we propose to learn a sparse + low-rank decomposition of the parameters of an Ising model using a convex regularized likelihood problem. We show that the same problem can be obtained as the dual of a maximum-entropy problem with a new type of relaxation, where the sample means collectively need to match the expected values only up to a given tolerance. The solution to the convex optimization problem has consistency properties in the high-dimensional setting, where the number of observed binary variables and the number of latent conditional Gaussian variables are allowed to grow with the number of training samples.

Keywords: Ising Models, Latent Variables, Sparse and Low-Rank Matrices, Maximum-Entropy Principle, High-Dimensional Consistency

1. Introduction

The principle of maximum entropy was proposed by Jaynes (1957) for probability density estimation. It states that from the probability densities that represent the current state of knowledge one should choose the one with the largest entropy, that is, the one which does not introduce additional biases. The state of knowledge is often given by sample points from a sample space and some fixed functions (sufficient statistics) on the sample space. The knowledge is then encoded naturally in form of constraints on the probability density by requiring that the expected values of the functions equal their respective sample means. Here, we assume the particularly simple multivariate sample space $\mathcal{X} = \{0,1\}^d$ and functions

$$\varphi_{ij} : x \mapsto x_i x_j \quad \text{for } i,j \in [d] = \{1,\ldots,d\}.$$

Suppose we are given sample points $x^{(1)}, \dots, x^{(n)} \in \mathcal{X}$. Then formally, for estimating the distribution from which the sample points are drawn, the principle of maximum entropy suggests solving the following entropy maximization problem

$$\max_{p \in \mathcal{P}} H(p) \quad \text{s.t. } \mathbb{E}[\varphi_{ij}] = \frac{1}{n} \sum_{k=1}^{n} \varphi_{ij}(x^{(k)}) \text{ for all } i, j \in [d],$$

where $\mathcal{P}$ is the set of all probability distributions on $\mathcal{X}$, the expectation is with respect to the distribution $p \in \mathcal{P}$, and $H(p) = -\mathbb{E}[\log p(x)]$ is the entropy. We denote the $(d \times d)$-matrix $\left(\frac{1}{n} \sum_{k=1}^{n} \varphi_{ij}(x^{(k)})\right)_{i,j \in [d]}$ of sample means compactly by Φ^n and the matrix of functions $\left(\varphi_{ij}\right)_{i,j \in [d]}$ by Φ. Then, the entropy maximization problem becomes

$$\max_{p \in \mathcal{P}} H(p) \quad \text{s.t. } \mathbb{E}[\Phi] - \Phi^n = 0.$$

Dudík et al. (2004) observed that invoking the principle of maximum entropy tends to overfit when the number of features d is large. Requiring that the expected values of the functions equal their respective sample means can be too restrictive. Consequently, they proposed to relax the constraint using the maximum norm as

$$\|\mathbb{E}[\Phi] - \Phi^n\|_\infty \leq c$$

for some $c > 0$. That is, for every function the expected value only needs to match the sample mean up to a tolerance of c. The dual of the relaxed problem has a natural interpretation as a feature-selective ℓ_1-regularized log-likelihood maximization problem

$$\max_{S \in \mathrm{Sym}(d)} \quad \ell(S) - c\|S\|_1,$$

where $\mathrm{Sym}(d)$ is the set of symmetric $(d \times d)$-matrices, $S \in \mathrm{Sym}(d)$ is the matrix of dual variables for the constraint $\|\mathbb{E}[\Phi] - \Phi^n\|_\infty \leq c$, and

$$\ell(S) = \langle S, \Phi^n \rangle - a(S)$$

is the log-likelihood function for pairwise Ising models with the standard matrix dot product $\langle S, \Phi^n \rangle = \mathrm{tr}\left(S^\top \Phi^n\right)$ and normalizer (log-partition function)

$$a(S) = \log \sum_{x \in \mathcal{X}} \langle S, \Phi(x) \rangle .$$

In this paper, we are restricting the relaxation of the entropy maximization problem by also enforcing the alternative constraint

$$\|\mathbb{E}[\Phi] - \Phi^n\| \leq \lambda,$$

where $\lambda > 0$ and $\|\cdot\|$ denotes the spectral norm on $\mathrm{Sym}(d)$. A difference to the maximum norm constraint is that now the expected values of the functions only need to collectively match the sample means up to a tolerance of λ instead of individually. The dual of the more strictly relaxed entropy maximization problem

$$\max_{p \in \mathcal{P}} H(p) \quad \text{s.t. } \|\mathbb{E}[\Phi] - \Phi^n\|_\infty \leq c \text{ and } \|\mathbb{E}[\Phi] - \Phi^n\| \leq \lambda$$

is the regularized log-likelihood maximization problem

$$\max_{S,\, L_1,\, L_2 \in \mathrm{Sym}(d)} \ell(S + L_1 - L_2) - c\|S\|_1 - \lambda \operatorname{tr}(L_1 + L_2) \quad \text{s.t. } L_1, L_2 \succeq 0,$$

see Appendix A in the full version Nussbaum and Giesen (2019) of this paper. Here, the regularization term $\operatorname{tr}(L_1 + L_2)$ promotes a low rank of the positive-semidefinite matrix $L_1 + L_2$. This implies that the matrix $L_1 - L_2$ in the log-likelihood function also has low rank. Thus, a solution of the dual problem is the sum of a sparse matrix S and a low-rank matrix $L_1 - L_2$. This can be interpreted as follows: the variables interact indirectly through the low-rank matrix $L_1 - L_2$, while some of the direct interactions through the matrix S are turned off by setting entries in S to zero. We get a more intuitive interpretation of the dual problem if we consider a weakening of the spectral norm constraint. The spectral norm constraint is equivalent to the two constraints

$$\mathbb{E}[\Phi] - \Phi^n \preceq \lambda \operatorname{Id} \quad \text{and} \quad \Phi^n - \mathbb{E}[\Phi] \preceq \lambda \operatorname{Id}$$

that bound the spectrum of the matrix $\mathbb{E}[\Phi] - \Phi^n$ from above and below. If we replace the spectral norm constraint by only the second of these two constraints in the maximum-entropy problem, then the dual problem becomes

$$\max_{S,\, L \in \mathrm{Sym}(d)} \ell(S + L) - c\|S\|_1 - \lambda \operatorname{tr}(L) \quad \text{s.t. } L \succeq 0.$$

This problem also arises as the log-likelihood maximization problem for a conditional Gaussian model (see Lauritzen (1996)) that exhibits observed binary variables and unobserved, latent conditional Gaussian variables. The sample space of the full mixed model is $\mathcal{X} \times \mathcal{Y} = \{0,1\}^d \times \mathbb{R}^l$, where $\mathcal{Y} = \mathbb{R}^l$ is the sample space for the unobserved variables. We want to write down the density of the conditional Gaussian model on this sample space. For that we respectively denote the interaction parameters between the observed binary variables by $S \in \mathrm{Sym}(d)$, the ones between the observed binary and latent conditional Gaussian variables by $R \in \mathbb{R}^{l \times d}$, and the ones between the latent conditional Gaussian variables by $\Lambda \in \mathrm{Sym}(l)$, where $\Lambda \succ 0$. Then, for $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and up to normalization, the density of the conditional Gaussian model is given as

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

One can check, see also Lauritzen (1996), that the conditional densities $p(y \mid x)$ are l-variate Gaussians on $\mathcal{Y}$. Here, we are interested in the marginal distribution

$$p(x) \propto \exp\left(\left\langle S + \frac{1}{2} R^\top \Lambda^{-1} R, \Phi(x) \right\rangle\right)$$

on $\mathcal{X}$ that is obtained by integrating over the unobserved variables in $\mathcal{Y}$, see Appendix B in the full version Nussbaum and Giesen (2019). The matrix $L = \frac{1}{2} R^\top \Lambda^{-1} R$ is symmetric and positive semidefinite. The log-likelihood function for the marginal model and the given data is thus given as

$$\ell(S + L) = \langle S + L, \Phi^n \rangle - a(S + L),$$

where $S, L \in \mathrm{Sym}(d)$, $L \succeq 0$ and $a(S + L)$ is once again the normalizer of the density.

If only a few of the binary variables interact directly, then S is sparse, and if the number of unobserved variables l is small compared to d, then L is of low rank. Hence, one could attempt to recover S and L from the data using the regularized log-likelihood maximization problem

$$\max_{S,\, L \in \mathrm{Sym}(d)} \ell(S+L) - c\|S\|_1 - \lambda \operatorname{tr}(L) \quad \text{s.t. } L \succeq 0 \tag{ML}$$

that we encountered before.

We are now in a similar situation as has been discussed by Chandrasekaran et al. (2012) who studied Gaussian graphical models with latent Gaussian variables. They were able to consistently estimate both the number of latent components, in our case l, and the conditional graphical model structure among the observed variables, in our case the zeroes in S. Their result holds in the high-dimensional setting, where the number of variables (latent and observed) may grow with the number of observed sample points. Here, we show a similar result for the Ising model with latent conditional Gaussian variables, that is, the one that we have introduced above.

2. Related Work

Graphical Models. The introduction of decomposed sparse + low-rank models followed a period of quite extensive research on sparse graphical models in various settings, for example Gaussians (Meinshausen and Bühlmann (2006), Ravikumar et al. (2011)), Ising models (Ravikumar et al. (2010)), discrete models (Jalali et al. (2011)), and more general conditional Gaussian and exponential family models (Lee and Hastie (2015), Lee et al. (2015), Cheng et al. (2017)). All estimators of sparse graphical models maximize some likelihood including a ℓ_1-penalty that induces sparsity.

Most of the referenced works contain high-dimensional consistency analyses that particularly aim at the recovery of the true graph structure, that is, the information which variables are *not* conditionally independent and thus interact. A prominent proof technique used throughout is the primal-dual-witness method originally introduced in Wainwright (2009) for the LASSO, that is, sparse regression. Generally, the assumptions necessary in order to be able to successfully identify the true interactions for graphical models (or rather the active predictors for the LASSO) are very similar. For example, one of the conditions that occurs repeatedly is irrepresentability, sometimes also referred to as incoherence. Intuitively, this condition limits the influence the active terms (edges) can have on the inactive terms (non-edges), see Ravikumar et al. (2011).

Sparse + low-rank models. The seminal work of Chandrasekaran et al. (2012) is the first to propose learning sparse + low-rank decompositions as an extension of classical graphical models. As such it has received a lot of attention since then, putting forth various commentators, for example Candès and Soltanolkotabi (2012), Lauritzen and Meinshausen (2012), and Wainwright (2012). Notably, Chandrasekaran et al. (2012)'s high-dimensional consistency analysis generalizes the proof-technique previously employed in graphical models. Hence, unsurprisingly, one of their central assumptions is a generalization of the irrepresentability condition.

Astoundingly, not so much effort has been undertaken in generalizing sparse + low-rank models to broader domains of variables. The particular case of multivariate binary models featuring a sparse + low-rank decomposition is related to Item Response Theory (IRT, see for example Hambleton et al. (1991)). In IRT the observed binary variables (test items) are usually assumed to be conditionally independent given some continuous latent variable (trait of the test taker). Chen et al. (2018) argued

that measuring conditional dependence by means of sparse + low-rank models might improve results from classical IRT. They estimate their models using pseudo-likelihood, a strategy that they also proposed in an earlier work, see Chen et al. (2016).

Chen et al. (2016) show that their estimator recovers the algebraic structure, that is, the conditional graph structure and the number of latent variables, with probability tending to one. However, their analysis only allows a growing number of sample points whereas they keep the number of variables fixed. Their result thus severs from the tradition to analyze the more challenging high-dimensional setting, where the number of variables is also explicitly tracked.

Placement of our work. Our main contribution is a high-dimensional consistency analysis of a likelihood estimator for multivariate binary sparse + low-rank models. Furthermore, our analysis is the first to show parametric consistency of the likelihood-estimates and to provide explicit rates for this type of models. It thus complements the existing literature. Our other contribution is the connection to a particular type of relaxed maximum-entropy problems that we established in the introduction. We have shown that this type of relaxation leads to an interpretation as the marginal model of a conditional Gaussian distribution. Interestingly, this has not drawn attention before, though our semidefiniteness constraints can be obtained as special cases of the general relaxed maximum-entropy problem discussed in Dudík and Schapire (2006).

3. Parametric and Algebraic Consistency

This section constitutes the main part of this paper. Here, we discuss assumptions that lead to consistency properties of the solution to the likelihood problem ML and state our consistency result. We are interested in the high-dimensional setting, where the number of samples n, the number of observed binary variables d, and the number of latent conditional Gaussian variables l are allowed to grow simultaneously. Meanwhile, there are some other problem-specific quantities that concern the curvature of the problem that we assume to be fixed. Hence, we keep the geometry of the problem fixed.

For studying the consistency properties, we use a slight reformulation of Problem ML from the introduction. First, we switch from a maximization to a minimization problem, and let ℓ be the *negative* log-likelihood from now on. Furthermore, we change the representation of the regularization parameters, namely

$$\begin{aligned}(S_n, L_n) \;=\; \underset{S,\,L}{\operatorname{argmin}} \quad & \ell(S+L) + \lambda_n\left(\gamma\|S\|_1 + \operatorname{tr} L\right) \\ \text{s.t.} \quad & L \succeq 0,\end{aligned} \tag{SL}$$

where γ controls the trade-off between the two regularization terms and λ_n controls the trade-off between the negative log-likelihood term and the regularization terms.

We want to point out that our consistency proof follows the lines of the seminal work in Chandrasekaran et al. (2012) who investigate a convex optimization problem for the parameter estimation of a model with observed and latent Gaussian variables. The main difference to the Ising model is that the Gaussian case requires a positive-definiteness constraint on the pairwise interaction parameter matrix $S + L$ that is necessary for normalizing the density. Furthermore, in the Gaussian case the pairwise interaction parameter matrix $S + L$ is the inverse of the covariance matrix. This is no longer the case for the Ising model, see Loh and Wainwright (2012).

In this work, we want to answer the question if it is possible to recover the parameters from data that has been drawn from a hypothetical *true* model distribution parametrized by $S^\star$ and $L^\star$. We focus on two key concepts of successful recovery in an asymptotic sense with high probability. The first is *parametric* consistency. This means that (S_n, L_n) should be close to $(S^\star, L^\star)$ w.r.t. some norm. Since the regularizer is the composed norm $\gamma\|S\|_1 + \operatorname{tr} L$, a natural norm for establishing parametric consistency is its dual norm

$$\|(S, L)\|_\gamma = \max\left\{\frac{\|S\|_\infty}{\gamma}, \|L\|\right\}.$$

The second type of consistency that we study is *algebraic* consistency. It holds if S_n recovers the true sparse support of $S^\star$, and if L_n has the same rank as $L^\star$.

In the following we discuss the assumptions for our consistency result. For that we proceed as follows: First, we discuss the requirements for parametric consistency of the compound matrix in Section 3.1. Next, we work out the three central assumptions that are sufficient for individual recovery of $S^\star$ and $L^\star$ in Section 3.2. We state our consistency result in Section 3.3. Finally, in Section 3.4 we outline the proof, the details of which can be found in the full version of this paper, see Nussbaum and Giesen (2019).

3.1. Parametric consistency of the compound matrix

In this section, we briefly sketch how the negative log-likelihood part of the objective function in Problem SL drives the compound matrix $\Theta_n = S_n + L_n$ that is constructed from the solution (S_n, L_n) to parametric consistency with high probability. We only consider the negative log-likelihood part because we assume that the relative weight λ_n of the regularization terms in the objective function goes to zero as the number of sample points goes to infinity. This implies that the estimated compound matrix is not affected much by the regularization terms since they contribute mostly small (but important) adjustments. More specifically, the ℓ_1-norm regularization on S shrinks entries of S such that entries of small magnitude are driven to zero such that S_n will likely be a sparse matrix. Likewise, the trace norm (or nuclear norm) can be thought of diminishing the singular values of the matrix L such that small singular values become zero, that is, L_n will likely be a low-rank matrix.

The negative log-likelihood function is strictly convex and thus has a unique minimizer $\hat{\Theta}$. We can assume that $\hat{\Theta} \approx \Theta_n$. Let $\Theta^\star = S^\star + L^\star$ and $\Delta_\Theta = \hat{\Theta} - \Theta^\star$. Then, consistent recovery of the compound matrix $\Theta^\star$ is essentially equivalent to the estimation error Δ_Θ being small. Now, consider the Taylor expansion

$$\ell(\Theta^\star + \Delta_\Theta) = \ell(\Theta^\star) + \nabla\ell(\Theta^\star)^\top\Delta_\Theta + \frac{1}{2}\Delta_\Theta^\top\nabla^2\ell(\Theta^\star)\Delta_\Theta + R(\Delta_\Theta)$$

with remainder $R(\Delta_\Theta)$. It turns out that if the number of samples is sufficiently large, then the gradient $\nabla\ell(\Theta^\star)$ is small with high probability, and if Δ_Θ is small, then the remainder $R(\Delta_\Theta)$ is also small. In this case, the Taylor expansion implies that locally around the true parameters the negative log-likelihood is well approximated by the quadratic form induced by its Hessian, namely

$$\ell(\Theta^\star + \Delta_\Theta) \approx \ell(\Theta^\star) + \frac{1}{2}\Delta_\Theta^\top\nabla^2\ell(\Theta^\star)\Delta_\Theta.$$

This quadratic form is obviously minimized at $\Delta_\Theta = 0$, which would entail consistent recovery of $\Theta^\star$ in a parametric sense. However, this does not explain how the sparse and low-rank components of $\Theta^\star$ can be recovered consistently. In the next section we elaborate sufficient assumptions for the consistent recovery of these components.

3.2. Assumptions for individual recovery

Consistent recovery of the components, more specifically parametric consistency of the solutions S_n and L_n, requires the two errors $\Delta_S = S_n - S^\star$ and $\Delta_L = L_n - L^\star$ to be small (in their respective norms). Both errors together form the joint error $\Delta_S + \Delta_L = \Theta_n - \Theta^\star \approx \Delta_\Theta$. Note though that the minimum of the quadratic form from the previous section at $\Delta_\Theta = 0$ does not imply that the individual errors Δ_S and Δ_L are small. We can only hope for parametric consistency of S_n and L_n if they are the unique solutions to Problem SL.

For uniqueness of the solutions we need to study optimality conditions. Problem SL is the Lagrange form of the constrained problem

$$\min\ \ell(S+L) \quad \text{s.t. } \|S\|_1 \le c_n \text{ and } \|L\|_* \le t_n$$

for suitable regularization parameters c_n and t_n, where we have neglected the positive-semidefiniteness constraint on L. The constraints can be thought of as convex relaxations of constraints that require S to have a certain sparsity and require L to have at most a certain rank. That is, S should be contained in the set of symmetric matrices of a given sparsity and L should be contained in the set of symmetric low-rank matrices. To formalize these sets we briefly review the varieties of sparse and low-rank matrices.

Sparse matrix variety. For $M \in \mathrm{Sym}(d)$ the support is defined as

$$\mathrm{supp}(M) = \{(i,j) \in [d] \times [d] :\ M_{ij} \neq 0\},$$

and the variety of sparse symmetric matrices with at most s non-zero entries is given as

$$\mathcal{S}(s) = \{S \in \mathrm{Sym}(d) : |\mathrm{supp}(S)| \le s\}.$$

Any matrix S with $|\mathrm{supp}(S)| = s$ is a smooth point of $\mathcal{S}(s)$ with tangent space

$$\Omega(S) = \{M \in \mathrm{Sym}(d) :\ \mathrm{supp}(M) \subseteq \mathrm{supp}(S)\}.$$

Low-rank matrix variety. The variety of matrices with rank at most r is given as

$$\mathcal{L}(r) = \{L \in \mathrm{Sym}(d) :\ \mathrm{rank}(L) \le r\}.$$

Any matrix L with rank r is a smooth point of $\mathcal{L}(r)$ with tangent space

$$T(L) = \left\{UX^\top + XU^\top :\ X \in \mathbb{R}^{d\times r}\right\},$$

where $L = UDU^\top$ is the restricted eigenvalue decomposition of L, that is, $U \in \mathbb{R}^{d\times r}$ has orthonormal columns and $D \in \mathbb{R}^{r\times r}$ is diagonal.

Next, we formulate conditions that ensure uniqueness in terms of the tangent spaces of the introduced varieties.

Transversality. Remember that we understand the constraints in the constrained formulation of Problem SL as convex relaxations of constraints of the form $S \in \mathcal{S}(s)$ and $L \in \mathcal{L}(r)$. Because the negative log-likelihood function ℓ is a function of $S + L$, its gradient with respect to S and its gradient with respect to L coincide at $S + L$. Hence, the first-order optimality conditions for the non-convex problem require that the gradient of the negative log-likelihood function needs to be normal to $\mathcal{S}(s)$ and $\mathcal{L}(r)$ at any (locally) optimal solutions $\hat{S}$ and $\hat{L}$, respectively. If the solution $(\hat{S}, \hat{L})$ is not (locally) unique, then basically the only way to get an alternative optimal solution that violates (local) uniqueness is by translating $\hat{S}$ and $\hat{L}$ by an element that is tangential to $\mathcal{S}(s)$ at $\hat{S}$ and tangential to $\mathcal{L}(r)$ at $\hat{L}$, respectively. Thus, it is necessary for (local) uniqueness of the optimal solution that such a tangential direction does not exist. Hence, the tangent spaces $\Omega(\hat{S})$ and $T(\hat{L})$ need to be *transverse*, that is, $\Omega(\hat{S}) \cap T(\hat{L}) = \{0\}$. Intuitively, if we require that transversality holds for the true parameters $(S^\star, L^\star)$, that is, $\Omega(S^\star) \cap T(L^\star) = \{0\}$, then provided that $(\hat{S}, \hat{L})$ is close to $(S^\star, L^\star)$, the tangent spaces $\Omega(\hat{S})$ and $T(\hat{L})$ should also be transverse.

We do not require transversality explicitly since it is implied by stronger assumptions that we motivate and state in the following. In particular, we want the (locally) optimal solutions $\hat{S}$ and $\hat{L}$ not only to be unique, but also to be *stable* under perturbations. This stability needs some additional concepts and notation that we introduce now.

Stability assumption. Here, stability means that if we perturb $\hat{S}$ and $\hat{L}$ in the respective tangential directions, then the gradient of the negative log-likelihood function should be far from being normal to the sparse and low-rank matrix varieties at the perturbed $\hat{S}$ and $\hat{L}$, respectively. As for transversality, we require stability for the true solution $(S^\star, L^\star)$ and expect that it carries over to the optimal solutions $\hat{S}$ and $\hat{L}$, provided they are close. More formally, we consider perturbations of $S^\star$ in directions from the tangent space $\Omega = \Omega(S^\star)$, and perturbations of $L^\star$ in directions from tangent spaces to the low-rank variety that are close to the true one $T = T(L^\star)$. The reason for considering tangent spaces close to $T(L^\star)$ is that there are low-rank matrices close to $L^\star$ that are not contained in $T(L^\star)$ because the low-rank matrix variety is locally curved at any smooth point.

Now, in light of a Taylor expansion the change of the gradient is locally governed by the data-independent Hessian $H^\star = \nabla^2 \ell(\Theta^\star) = \nabla^2 a(\Theta^\star)$ of the negative log-likelihood function at $\Theta^\star$. To make sure that the gradient of the tangentially perturbed (true) solution cannot be normal to the respective matrix varieties we require that it has a significant component in the tangent spaces at the perturbed solution. This is achieved if the *minimum gains* of the Hessian $H^\star$ in the respective tangential directions

$$\alpha_\Omega = \min_{M \in \Omega,\, \|M\|_\infty = 1} \|P_\Omega H^\star M\|_\infty, \quad \text{and}$$
$$\alpha_{T,\varepsilon} = \min_{\rho(T,T') \leq \varepsilon} \; \min_{M \in T',\, \|M\| = 1} \|P_{T'} H^\star M\|$$

are large, where $T' \subseteq \mathrm{Sym}(d)$ are tangent spaces to the low-rank matrix variety that are close to T in terms of the *twisting*

$$\rho(T, T') = \max_{\|M\| = 1} \|[P_T - P_{T'}](M)\|$$

between these subspaces given some $\varepsilon > 0$. Here, we denote projections onto a matrix subspace by P subindexed by the subspace.

Note though that only requiring α_Ω and $\alpha_{T,\varepsilon}$ to be large is not enough if the *maximum effects* of the Hessian $H^\star$ in the respective normal directions

$$\delta_\Omega = \max_{M \in \Omega,\, \|M\|_\infty = 1} \|P_{\Omega^\perp} H^\star M\|_\infty, \quad \text{and}$$
$$\delta_{T,\varepsilon} = \max_{\rho(T,T') \leq \varepsilon} \max_{M \in T',\, \|M\| = 1} \|P_{T'^\perp} H^\star M\|$$

are also large, because then the gradient of the negative log-likelihood function at the perturbed (true) solution could still be almost normal to the respective varieties. Here, $\Omega^\perp$ is the normal space at $S^\star$ orthogonal to Ω, and $T'^\perp$ is the space orthogonal to T'.

Overall, we require that $\alpha_\varepsilon = \min\{\alpha_\Omega, \alpha_{T,\varepsilon}\}$ is bounded away from zero and that the ratio $\delta_\varepsilon / \alpha_\varepsilon$ is bounded from above, where $\delta_\varepsilon = \max\{\delta_\Omega, \delta_{T,\varepsilon}\}$. Note that in our definitions of the minimum gains and maximum effects we used the ℓ_∞- and the spectral norm, which are dual to the ℓ_1- and the nuclear norm, respectively. Ultimately, we want to express the stability assumption in the $\|\cdot\|_\gamma$-norm which is the dual norm to the regularization term in Problem SL. For that we need to compare the ℓ_∞- and the spectral norm. This can be accomplished by using norm compatibility constants that are given as the smallest possible $\xi(T(L))$ and $\mu(\Omega(S))$ such that

$$\|M\|_\infty \leq \xi(T(L))\|M\| \text{ for all } M \in T(L), \text{ and } \|N\| \leq \mu(\Omega(S))\|N\|_\infty \text{ for all } N \in \Omega(S),$$

where $\Omega(S)$ and $T(L)$ are the tangent spaces at points S and L from the sparse matrix variety $\mathcal{S}(|\operatorname{supp} S|)$ and the low-rank matrix variety $\mathcal{L}(\operatorname{rank} L)$, respectively. Let us now specify our assumptions in terms of the stability constants from above.

Assumption 1 (Stability) *We set $\varepsilon = \xi(T)/2$ and assume that*

1. *$\alpha = \alpha_{\xi(T)/2} > 0$, and*
2. *there exists $\nu \in (0, \frac{1}{2}]$ such that $\frac{\delta}{\alpha} \leq 1 - 2\nu$, where $\delta = \delta_{\xi(T)/2}$.*

The second assumption is essentially a generalization of the well-known irrepresentability condition, see for example Ravikumar et al. (2011). The next assumption ensures that there are values of γ for which stability can be expressed in terms of the $\|\cdot\|_\gamma$-norm, that is, a coupled version of stability.

γ-feasibility assumption. The norm compatibility constants $\mu(\Omega)$ and $\xi(T)$ allow further insights into the realm of problems for which consistent recovery is possible. First, it can be shown, see Chandrasekaran et al. (2011), that $\mu(\Omega) \leq \deg_{\max}(S^\star)$, where $\deg_{\max}(S^\star)$ is the maximum number of non-zero entries per row/column of $S^\star$, that is, $\mu(\Omega)$ constitutes a lower bound for $\deg_{\max}(S^\star)$. Intuitively, if $\deg_{\max}(S^\star)$ is large, then the non-zero entries of the sparse matrix $S^\star$ could be concentrated in just a few rows/columns and thus $S^\star$ would be of low rank. Hence, in order not to confuse $S^\star$ with a low-rank matrix we want the lower bound $\mu(\Omega)$ on the maximum degree $\deg_{\max}(S^\star)$ to be small.

Second, $\xi(T)$ constitutes a lower bound on the *incoherence* of the matrix $L^\star$. Incoherence measures how well a subspace is aligned with the standard coordinate axes. Formally, the incoherence of a

subspace $U \subset \mathbb{R}^d$ is defined as $\mathrm{coh}(U) = \max_i \|P_U e_i\|$ where the e_i are the standard basis vectors of $\mathbb{R}^d$. It is known, see again Chandrasekaran et al. (2011), that

$$\xi(T) = \xi(T(L^\star)) \leq 2\,\mathrm{coh}(L^\star),$$

where $\mathrm{coh}(L^\star)$ is the incoherence of the subspace spanned by the rows/columns of the symmetric matrix $L^\star$. A large value $\mathrm{coh}(L^\star)$ means that the row/column space of $L^\star$ is well aligned with the standard coordinate axes. In this case, the entries of $L^\star$ do not need to be spread out and thus $L^\star$ could have many zero entries, that is, it could be a sparse matrix. Hence, in order not to confuse $L^\star$ with a sparse matrix we want the lower bound $\xi(T)/2$ on the incoherence $\mathrm{coh}(L^\star)$, or equivalently $\xi(T)$, to be small.

Altogether, we want both $\mu(\Omega)$ and $\xi(T)$ to be small to avoid confusion of the sparse and the low-rank parts. Now, in Problem SL, the parameter $\gamma > 0$ controls the trade-off between the regularization term that promotes sparsity, that is, the ℓ_1-norm term, and the regularization term that promotes low rank, that is, the nuclear norm term. It turns out that the range of values for γ that are feasible for our consistency analysis becomes larger if $\mu(\Omega)$ and $\xi(T)$ are small. Indeed, the following assumption ensures that the range of values of γ that are feasible for our consistency analysis is non-empty.

Assumption 2 (γ-feasibility) *The range $[\gamma_{\min}, \gamma_{\max}]$ with*

$$\gamma_{\min} = \frac{3\beta(2-\nu)\xi(T)}{\nu\alpha} \quad \textit{and} \quad \gamma_{\max} = \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}.$$

is non-empty. Here, we use the additional problem-specific constant $\beta = \max\{\beta_\Omega, \beta_T\}$ with

$$\beta_\Omega = \max_{M \in \Omega,\, \|M\|=1} \|H^\star M\|, \quad \textit{and}$$
$$\beta_T = \max_{\rho(T,T') \leq \frac{\xi(T)}{2}} \; \max_{M \in T',\, \|M\|_\infty = 1} \|H^\star M\|_\infty.$$

The γ-feasibility assumption is equivalent to

$$\mu(\Omega)\xi(T) \leq \frac{1}{6}\left(\frac{\nu\alpha}{\beta(2-\nu)}\right)^2.$$

Note that this upper bound on the product $\mu(\Omega)\xi(T)$ is essentially controlled by the product $\nu\alpha$. It is easier to satisfy when the latter product is large. This is well aligned with the stability assumption, because in terms of the stability assumption the good case is that the product $\nu\alpha$ is large, or more specifically that α is large and ν is close to $1/2$.

Gap assumption. Intuitively, if the smallest-magnitude non-zero entry $s_{\min}$ of $S^\star$ is too small, then it is difficult to recover the support of $S^\star$. Similarly, if the smallest non-zero eigenvalue $\sigma_{\min}$ of $L^\star$ is too small, then it is difficult to recover the rank of $L^\star$. Hence, we make the following final assumption.

Assumption 3 (Gap) *We require that*

$$s_{\min} \geq \frac{C_S \lambda_n}{\mu(\Omega)} \quad \textit{and} \quad \sigma_{\min} \geq \frac{C_L \lambda_n}{\xi(T)^2},$$

where C_S and C_L are problem-specific constants that are specified more precisely later.

Recall that the regularization parameter λ_n controls how strongly the eigenvalues of the solution L_n and the entries of the solution S_n are driven to zero. Hence, the required gaps get weaker as the number of sample points grows, because the parameter λ_n goes to zero as n goes to infinity.

3.3. Consistency theorem

We state our consistency result using problem-specific data-independent constants C_1, C_2 and C_3. Their exact definitions can be found alongside the proof in (Nussbaum and Giesen, 2019, Section 5), that is, in the full version of this paper. Also note that the norm compatibility constant $\xi(T)$ is implicitly related to the number of latent variables l. This is because $\xi(T) \leq 2\,\mathrm{coh}(L^\star)$ as we have seen above and $\sqrt{l/d} \leq \mathrm{coh}(L^\star) \leq 1$, see Chandrasekaran et al. (2011). Hence, the smaller l, the better can the upper bound on $\xi(T)$ be. Therefore, we track $\xi(T)$ and $\mu(\Omega)$ explicitly in our analysis.

Theorem 1 (Consistency) *Let $S^\star \in Sym(d)$ be a sparse and let $0 \preceq L^\star \in Sym(d)$ be a low-rank matrix. Denote by $\Omega = \Omega(S^\star)$ and $T = T(L^\star)$ the tangent spaces at $S^\star$ and $L^\star$, respectively to the variety of symmetric sparse matrices and to the variety of symmetric low-rank matrices. Suppose that we observed samples $x^{(1)}, \ldots, x^{(n)}$ drawn from a pairwise Ising model with interaction matrix $S^\star + L^\star$ such that the stability assumption, the γ-feasibility assumption, and the gap assumption hold. Moreover let $\kappa > 0$, and assume that for the number of sample points n it holds that*

$$n > \frac{C_1 \kappa}{\xi(T)^4} d \log d,$$

and that the regularization parameter λ_n it set as

$$\lambda_n = \frac{C_2}{\xi(T)} \sqrt{\frac{\kappa d \log d}{n}}.$$

Then, it follows with probability at least $1 - d^{-\kappa}$ that the solution (S_n, L_n) to the convex program SL is

a) parametrically consistent, that is, $\|(S_n - S^\star, L_n - L^\star)\|_\gamma \leq C_3 \lambda_n$, and

b) algebraically consistent, that is, S_n and $S^\star$ have the same support (actually, the signs of corresponding entries coincide), and L_n and $L^\star$ have the same ranks.

3.4. Outline of the proof

The proof of Theorem 1 is similar to the one given in Chandrasekaran et al. (2012) for latent variable models with observed Gaussians. More generally, it builds on a version of the primal-dual-witness proof technique. The proof consists of the following main steps:

(1) First, we consider the *correct model set* $\mathcal{M}$ whose elements are all parametrically and algebraically consistent under the stability, γ-feasibility, and gap assumptions. Hence, any solution $(S_\mathcal{M}, L_\mathcal{M})$ to our problem, if additionally constrained to $\mathcal{M}$, is consistent.

(2) Second, since the set $\mathcal{M}$ is non-convex, we consider a simplified and linearized version $\mathcal{Y}$ of the set $\mathcal{M}$ and show that the solution $(S_\mathcal{Y}, L_\mathcal{Y})$ to the problem constrained to the linearized model space $\mathcal{Y}$ is unique and equals $(S_\mathcal{M}, L_\mathcal{M})$. Since it is the same solution, consistency follows from the first step.

(3) Third, we show that the solution $(S_{\mathcal{M}}, L_{\mathcal{M}}) = (S_{\mathcal{Y}}, L_{\mathcal{Y}})$ also solves Problem SL. More precisely, we show that this solution is *strictly dual feasible* and hence can be used as a witness as required for the primal-dual-witness technique. This implies that it is also the unique solution, with all the consistency properties from the previous steps.

(4) Finally, we show that the assumptions from Theorem 1 entail all those made in the previous steps with high probability. Thereby, the proof is concluded.

4. Discussion

Our result, that constitutes the first high-dimensional consistency analysis for sparse + low-rank Ising models, requires slightly more samples (in the sense of an additional logarithmic factor $\log d$, and polynomial probability) than were required for consistent recovery for the sparse + low-rank Gaussian models considered by Chandrasekaran et al. (2012). This is because the strong tail properties of multivariate Gaussian distributions do not hold for multivariate Ising distributions. Hence, it is more difficult to bound the sampling error $\mathbb{E}[\Phi] - \Phi^n$ of the second-moment matrices, which results in weaker probabilistic spectral norm bounds of this sampling error. Under our assumptions, we believe that the sampling complexity, that is, the number of samples required for consistent recovery of sparse + low-rank Ising models, cannot be improved. We also provided a detailed discussion of why all of our assumptions are important.

Acknowledgments

We gratefully acknowledge financial support from the German Science Foundation (DFG) grant (GI-711/5-1) within the priority program (SPP 1736) Algorithms for Big Data.

References

Emmanuel J. Candès and Mahdi Soltanolkotabi. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1996–2004, 2012.

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.

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.

Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. A fused latent and graphical model for multivariate binary data. Technical report, arXiv preprint arXiv:1606.08925, 2016.

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.

Jie Cheng, Tianxi Li, Elizaveta Levina, and Ji Zhu. High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26(2):367–378, 2017.

Miroslav Dudík and Robert E. Schapire. Maximum entropy distribution estimation with generalized regularization. In *Conference on Learning Theory (COLT)*, pages 123–138, 2006.

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Conference on Learning Theory (COLT)*, pages 472–486, 2004.

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

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.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

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

Steffen L. Lauritzen and Nicolai Meinshausen. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1973–1977, 2012.

Jason D. Lee and Trevor J. Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

Jason D. Lee, Yuekai Sun, and Jonathan E. Taylor. On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics*, 9(1):608–642, 2015.

Po-Ling Loh and Martin J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. In *Conference on Neural Information Processing Systems (NIPS)*, pages 2096–2104, 2012.

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

Frank Nussbaum and Joachim Giesen. Ising models with latent conditional Gaussian variables. Technical report, arXiv preprint arXiv:1901.09712, 2019.

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.

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.

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.

Martin J. Wainwright. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1978–1983, 2012.