Ising Models with Latent Continuous Variables

Joachim Giesen (joint work with Frank Nussbaum)

Friedrich-Schiller-Universität Jena

Ising models

Ising models are probability distributions on the sample space

$$\mathcal{X} = \{0, 1\}^n$$

of the form

$$p(x) \propto \exp\left(x^\top S x\right), \quad S \in \mathrm{Sym}(n).$$

Ising models

Ising models are probability distributions on the sample space

$$\mathcal{X} = \{0, 1\}^n$$

of the form

$$p(x) \propto \exp\left(x^\top S x\right), \quad S \in \mathrm{Sym}(n).$$

Typically only a few of the variables interact with each other, thus S is sparse.

Ising models

Ising models are probability distributions on the sample space

$$\mathcal{X} = \{0, 1\}^n$$

of the form

$$p(x) \propto \exp\left(x^\top S x\right), \quad S \in \mathrm{Sym}(n).$$

Typically only a few of the variables interact with each other, thus S is sparse.

Ising models are also known in machine learning as Boltzmann machines.

Multivariate Gaussians

Multivariate Gaussians are probability distributions on the sample space

$$\mathcal{Y} = \mathbb{R}^m$$

of the form

$$p(y) \propto \exp\left(-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right), \quad \Sigma \in \mathrm{PD}(m).$$

Multivariate Gaussians

Multivariate Gaussians are probability distributions on the sample space

$$\mathcal{Y} = \mathbb{R}^m$$

of the form

$$p(y) \propto \exp\left(-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right), \quad \Sigma \in \mathrm{PD}(m).$$

Typically only a few of the variables interact with each other, and thus $\Lambda = \Sigma^{-1}$ is sparse (Gaussian graphical model).

CG distributions

Restricted CG distributions that are defined on the sample space

$$\mathcal{X} \times \mathcal{Y} = \{0, 1\}^n \times \mathbb{R}^m$$

are of the form

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

Ising models with latent continuous variables

Marginalizing out the continuous variables from a CG distribution gives the marginal distribution on

$$\mathcal{X} = \{0, 1\}^n$$

of the form

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

Ising models with latent continuous variables

Marginalizing out the continuous variables from a CG distribution gives the marginal distribution on

$$\mathcal{X} = \{0, 1\}^n$$

of the form

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

If $m \ll n$ (number of continuous variables is much smaller than the number Bernoulli variables), then the PSD matrix $L = \frac{1}{2}R^\top \Lambda R$ is of small rank.

Likelihood function for latent variable Ising model

Given data points

$$x^{(1)}, \dots, x^{(k)} \in \mathcal{X} = \{0, 1\}^n,$$

we want to estimate the model parameters $S \in \mathrm{Sym}(n)$ (sparse) and $L \in \mathrm{Sym}(n)$ (PSD and low rank).

Likelihood function for latent variable Ising model

Given data points

$$x^{(1)}, \dots, x^{(k)} \in \mathcal{X} = \{0, 1\}^n,$$

we want to estimate the model parameters $S \in \mathrm{Sym}(n)$ (sparse) and $L \in \mathrm{Sym}(n)$ (PSD and low rank).

The log-likelihood function for the latent variable Ising model is

$$\ell(S + L) = \sum_{i=1}^{n} {x^{(i)}}^{\top}(S + L)x^{(i)} - a(S + L),$$

where a is a normalization function.

Promoting sparse + low rank solutions

Regularized log-likelihood problem

$$\max_{S,L} \; \ell(S+L) - c\|S\|_1 - \lambda \mathrm{Tr}(L) \quad \text{s.t. } L \succeq 0,$$

where $c, \lambda > 0$ are regularization parameters.

Promoting sparse + low rank solutions

Regularized log-likelihood problem

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

where $c, \lambda > 0$ are regularization parameters.

Question, under which conditions can we recover S and L individually from a solution of the regularized log-likelihood problem?

Alternative derivation of the problem

Maximum Entropy Principle [Jaynes 1955]

From the probability distributions that represent the current state of knowledge choose the one with largest entropy.

Maximum entropy principle

Current state of knowledge:

1. Sample points $x^{(1)}, \ldots, x^{(k)}$ drawn from the sample space $\mathcal{X} = \{0, 1\}^n$.
2. Functions on $\mathcal{X}$ (sufficient statistics). Here we consider

$$\varphi_{ij} : x = (x_1, \ldots, x_n) \mapsto x_i x_j, \quad i, j \in [n].$$

Entropy maximization problem

$$\max_{p \in \mathcal{P}} \; H(p) \quad \text{s.t. } E[\varphi_{ij}] = \frac{1}{n} \sum_{l=1}^{k} \varphi_{ij}(x^{(l)})$$

Entropy maximization problem

$$\max_{p \in \mathcal{P}} \; H(p) \quad \text{s.t. } E[\varphi_{ij}] = \frac{1}{n} \sum_{l=1}^{k} \varphi_{ij}(x^{(l)})$$

Or more compactly

$$\max_{p \in \mathcal{P}} \; H(p) \quad \text{s.t. } E[\Phi] = \Phi^k,$$

if we collect the functions φ_{ij} in the $n \times n$ matrix Φ and set $\Phi^k = \sum_{l=1}^{k} x^{(l)} {x^{(l)}}^\top$.

Relaxed entropy maximization and its dual

The problem with relaxed constraint reads as

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

Relaxed entropy maximization and its dual

The problem with relaxed constraint reads as

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

whose Lagrangian dual is given as

$$\max_{S} \; \ell(S) - c\|S\|_1$$

with $S \in \mathrm{Sym}(n)$.

Relaxed entropy maximization and its dual

The problem with relaxed constraint reads as

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

whose Lagrangian dual is given as

$$\max_{S} \; \ell(S) - c\|S\|_1$$

with $S \in \mathrm{Sym}(n)$.

Maximum entropy – maximum likelihood duality

Spectral norm relaxation

Alternatively/additionally we can impose the constraint

$$\|E[\Phi] - \Phi^k\|_2 \leq \lambda,$$

Spectral norm relaxation

Alternatively/additionally we can impose the constraint

$$\|E[\Phi] - \Phi^k\|_2 \leq \lambda,$$

which gives

$$\begin{aligned} \max_{p \in \mathcal{P}} \;& H(p) \\ \text{s.t.} \;& \|E[\Phi] - \Phi^k\|_\infty \leq c, \\ & \|E[\Phi] - \Phi^k\|_2 \leq \lambda. \end{aligned}$$

Dual of spectral norm relaxed problem ...

... is the regularized maximum likelihood problem

$$\begin{aligned} \max_{S, L_1, L_2} \quad & \ell(S - L_1 + L_2) - c\|S\|_1 - \lambda \mathrm{Tr}(L_1 + L_2) \\ \text{s.t.} \quad & L_1, L_2 \succeq 0, \end{aligned}$$

where $S, L_1, L_2 \in \mathrm{Sym}(n)$.

Dual of spectral norm relaxed problem ...

... is the regularized maximum likelihood problem

$$\begin{aligned} \max_{S, L_1, L_2} \quad & \ell(S - L_1 + L_2) - c\|S\|_1 - \lambda \mathrm{Tr}(L_1 + L_2) \\ \text{s.t.} \quad & L_1, L_2 \succeq 0, \end{aligned}$$

where $S, L_1, L_2 \in \mathrm{Sym}(n)$.

The regularization term $\mathrm{Tr}(L_1 + L_2)$ promotes a low rank of $L_1 + L_2$, and thus also of $L_2 - L_1$.

Dual of spectral norm relaxed problem ...

... is the regularized maximum likelihood problem

$$\begin{aligned}
\max_{S, L_1, L_2} \quad & \ell(S - L_1 + L_2) - c\|S\|_1 - \lambda \mathrm{Tr}(L_1 + L_2) \\
\text{s.t.} \quad & L_1, L_2 \succeq 0,
\end{aligned}$$

where $S, L_1, L_2 \in \mathrm{Sym}(n)$.

The regularization term $\mathrm{Tr}(L_1 + L_2)$ promotes a low rank of $L_1 + L_2$, and thus also of $L_2 - L_1$.

Hence, the interaction matrix $S - L_1 + L_2$ has a

sparse (S) + *low rank* $(L_2 - L_1)$ *decomposition.*

Weakening of the spectral norm constraint

The spectral norm constraint

$$\|E[\Phi] - \Phi^k\|_2 \leq \lambda,$$

can also be written as

$$E[\Phi] - \Phi^k \preceq \lambda \quad \text{and} \quad \Phi^k - E[\Phi] \preceq \lambda$$

Weakening of the spectral norm constraint

The spectral norm constraint

$$\|E[\Phi] - \Phi^k\|_2 \leq \lambda,$$

can also be written as

$$E[\Phi] - \Phi^k \preceq \lambda \quad \text{and} \quad \Phi^k - E[\Phi] \preceq \lambda$$

Skipping the first constraint gives

$$\begin{aligned} \max_{p \in \mathcal{P}} \ & H(p) \\ \text{s.t.} \ & \|E[\Phi] - \Phi^k\|_\infty \leq c, \\ & \Phi^k - E[\Phi] \preceq \lambda. \end{aligned}$$

Weakening of the spectral norm constraint

The spectral norm constraint

$$\|E[\Phi] - \Phi^k\|_2 \leq \lambda,$$

can also be written as

$$E[\Phi] - \Phi^k \preceq \lambda \quad \text{and} \quad \Phi^k - E[\Phi] \preceq \lambda$$

Skipping the first constraint gives

$$\begin{aligned} \max_{p \in \mathcal{P}} \quad & H(p) \\ \text{s.t.} \quad & \|E[\Phi] - \Phi^k\|_\infty \leq c, \\ & \Phi^k - E[\Phi] \preceq \lambda. \end{aligned}$$

Whose dual is given as our marginal model

$$\max_{S,L} \; \ell(S + L) - c\|S\|_1 - \lambda \mathrm{Tr}(L) \quad \text{s.t.} \; L \succeq 0.$$

Consistency guarantees

We consider the slightly reformulated problem

$$\max_{S,L} \ \ell(S+L) - \lambda_k\big(\gamma\|S\|_1 + \mathrm{Tr}(L)\big) \quad \text{s.t. } L \succeq 0,$$

where the likelihood function ℓ depends on the sample points $x^{(1)}, \ldots, x^{(k)}$ through the covariance matrix Φ^k, and λ_k goes to zero with growing k.

Consistency guarantees

We consider the slightly reformulated problem

$$\max_{S,L} \; \ell(S+L) - \lambda_k\big(\gamma\|S\|_1 + \mathrm{Tr}(L)\big) \quad \text{s.t. } L \succeq 0,$$

where the likelihood function ℓ depends on the sample points $x^{(1)}, \ldots, x^{(k)}$ through the covariance matrix Φ^k, and λ_k goes to zero with growing k.

Assume that the sample points are drawn from distribution with interaction parameter $S^\star, L^\star \in \mathrm{Sym}(n)$, where $S^\star$ is sparse and $L^\star$ is of low rank.

Consistency guarantees

We consider the slightly reformulated problem

$$\max_{S,L} \ \ell(S+L) - \lambda_k\big(\gamma\|S\|_1 + \mathrm{Tr}(L)\big) \quad \text{s.t. } L \succeq 0,$$

where the likelihood function ℓ depends on the sample points $x^{(1)}, \ldots, x^{(k)}$ through the covariance matrix Φ^k, and λ_k goes to zero with growing k.

Assume that the sample points are drawn from distribution with interaction parameter $S^\star, L^\star \in \mathrm{Sym}(n)$, where $S^\star$ is sparse and $L^\star$ is of low rank.

1. Can we approximate $S^\star$ and $L^\star$ from a solution to the regularized likelihood problem?

Consistency guarantees

We consider the slightly reformulated problem

$$\max_{S,L} \; \ell(S+L) - \lambda_k\big(\gamma\|S\|_1 + \mathrm{Tr}(L)\big) \quad \text{s.t. } L \succeq 0,$$

where the likelihood function ℓ depends on the sample points $x^{(1)}, \ldots, x^{(k)}$ through the covariance matrix Φ^k, and λ_k goes to zero with growing k.

Assume that the sample points are drawn from distribution with interaction parameter $S^\star, L^\star \in \mathrm{Sym}(n)$, where $S^\star$ is sparse and $L^\star$ is of low rank.

1. Can we approximate $S^\star$ and $L^\star$ from a solution to the regularized likelihood problem?
2. Does the solution recover the sparsity of $S^\star$ and the rank of $L^\star$?

Problem: non-identifiability

The matrix

$$M = \begin{pmatrix} 1 & 0 & \dots \\ 0 & 0 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

is sparse and of low rank.

Problem: non-identifiability

The matrix

$$M = \begin{pmatrix} 1 & 0 & \dots \\ 0 & 0 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

is sparse and of low rank.

In general, we cannot distinguish $(S^\star, L^\star)$ from $(S^\star + M, L^\star - M)$, because both have the same compound matrix $S^\star + L^\star$.

Transversality assumption

Idea: Uniqueness of the solution of

$$\max_{S,L} \; \ell(S+L) - \lambda_k\big(\gamma\|S\|_1 + \mathrm{Tr}(L)\big)$$

is necessary for individual recovery.

Transversality assumption

Idea: Uniqueness of the solution of

$$\max_{S,L} \; \ell(S+L) - \lambda_k\big(\gamma\|S\|_1 + \mathrm{Tr}(L)\big)$$

is necessary for individual recovery.

Optimality can be geometrically characterized as:

> *The gradient $\nabla\ell(S,L)$ needs to be normal to the tangent space of the variety of sparse matrices at S and also normal to the tangent space of the variety of low rank matrices at L.*

Transversality assumption

Idea: Uniqueness of the solution of

$$\max_{S,L} \; \ell(S+L) - \lambda_k\big(\gamma\|S\|_1 + \mathrm{Tr}(L)\big)$$

is necessary for individual recovery.

Optimality can be geometrically characterized as:

> *The gradient $\nabla\ell(S,L)$ needs to be normal to the tangent space of the variety of sparse matrices at S and also normal to the tangent space of the variety of low rank matrices at L.*

For uniqueness we need that the two tangent spaces are *transversal*, i.e., only share the origin.

Gap assumption

We also need to require that

$$s_{\min} \geq c_S \lambda_k \quad \text{and} \quad \sigma_{\min} \geq c_L \lambda_k,$$

where $s_{\min}$ is the smallest magnitude of any non-zero entry in $S^\star$ and $\sigma_{\min}$ is the smallest non-zero eigenvalue of $L^\star$. Furthermore, c_S and c_L are positive constants.

Consistency theorem

Theorem Let $(S^\star, L^\star)$ be the true model parameters and (S_k, L_k) be the solution to the regularized likelihood problem. Let

$$k > c_1 \cdot t \cdot n \log n \quad \text{and} \quad \lambda_k = c_2 \sqrt{\frac{t \cdot n \log n}{k}}$$

for constants $c_1, c_2, c_3, t > 0$. Then with probability at least $1 - k^{-t}$

1. $\max\left\{\|S_k - S^\star\|_\infty, \|L_k - L^\star\|_2\right\} \leq c_3 \lambda_k$, and
2. S_k and $S^\star$ have the same support, and L_k and $L^\star$ have the same rank.

Consistency theorem

Theorem Let $(S^\star, L^\star)$ be the true model parameters and (S_k, L_k) be the solution to the regularized likelihood problem. Let

$$k > c_1 \cdot t \cdot n \log n \quad \text{and} \quad \lambda_k = c_2 \sqrt{\frac{t \cdot n \log n}{k}}$$

for constants $c_1, c_2, c_3, t > 0$. Then with probability at least $1 - k^{-t}$

1. $\max \left\{ \|S_k - S^\star\|_\infty, \|L_k - L^\star\|_2 \right\} \leq c_3 \lambda_k$, and
2. S_k and $S^\star$ have the same support, and L_k and $L^\star$ have the same rank.

Proof Similar to the consistency proof for Gaussian latent variable graphical models by Chandrasekaran, Parrilo and Willsky.