High dimensional ising model selection using $\ell_1$-regularized logistic regression

Jiawei Wen

$^1$Department of Statistics
Pennsylvania State University

597 Presentation 2016
Outline

1. Introduction
   - Background and Motivation
   - Problem Formulation

2. Method and theoretical guarantees
   - Theoretical results
   - Primal-dual witness for graph recovery
   - Analysis under sample Fisher matrix assumptions
   - Extensions to general discrete MRF.

3. Experimental results.

4. Conclusion and framework
Background

Markov random fields (MRF)/Undirected graphical model.

- MRF: undirected graph $G = (V, E)$ with vertex set $V = \{1, 2, \cdots, p\}$ and edge set $E \subset V \times V$
- Structure of the graph: encodes certain conditional independence assumptions among subsets of $X = (X_1, \cdots, X_p)$, where each $X_s$ is associated with a vertex $s \in V$.

One important problem:
Estimate the underlying graph from $n$ i.i.d samples $\{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$ drawn from the distribution specified by the MRF.
Previous work on MRF

- **Constraint-based approaches**: estimates conditional independence from the data using hypothesis testing.

- **Score-based approaches**: combine a metric for the complexity of the graph with a measure of the goodness of fit, e.g., log-likelihood of the maximum likelihood parameters.

**Drawbacks**: Calculation for the partition function associated with the MRF is computationally intractable.
Basic approach

- Basic approach: performing sequential $\ell_1$-regularized logistic regression. Then using the sparsity pattern of the regression vector to infer the underlying neighborhood structure.
- Sample size: $n$; model dimension $p$; maximum neighborhood size $d$. Both $p$ and $d$ may tend to infinity as a function of the size $n$.
- Focus on binary Ising model and generalize the method to general discrete MRF.
**Background**

- **Pairwise Markov random fields.**
  \[ X = (X_1, X_2, \cdots, X_p) \] is a random vector. \( G = (V, E) \) is an undirected graph with vertex set \( V = \{1, \cdots, p\} \) and edge set \( E \). The pairwise MRF is the family of distributions of \( X \):
  \[
P(x) \propto \exp\left\{ \sum_{(s,t) \in E} \phi_{st}(x_s, x_t) \right\}
  \]

- **Ising model.** We focus on the special case of the Ising model in which \( X_s \in \{-1, 1\} \) for each vertex \( s \in V \), and
  \[ \phi_{st}(x_s, x_t) = \theta^*_{st} x_s x_t, \quad \theta^*_{st} \in \mathcal{R}, \]
  thus
  \[
P(x) = \frac{1}{Z(\theta^*)} \exp \left\{ \sum_{(s,t) \in E} \theta^*_{st} x_s x_t \right\}.
  \]

\[ (1) \]
Graphical model selection

- **Goal.** \( \mathcal{X}_1^n := \{x^{(1)}, \ldots, x^{(n)}\} \) of \( n \) samples where \( x^{(i)} \in \{-1, +1\}^p \) is drawn i.i.d. from a distribution \( P_{\theta^*} \).

The goal of graphical model selection is to **infer the edge set** \( E \) (the same as estimate \( \theta_{st}^* \)).

- **Stronger criterion of signed edge recovery.**
  Given a graphical model with parameter \( \theta^* \), define the edge sign vector

\[
E^* = \begin{cases} 
\text{sign}(\theta_{st}^*), & \text{if } (s, t) \in E \\
0, & \text{otherwise.}
\end{cases}
\]  

(2)
• Classical notion of statistical consistency: limiting behavior of the estimator as $n \to \infty$ and $p$ fixed.

• Analysis in this paper is of high-dimensional nature. As $n \to \infty$, $p = p(n)$, $d = d(n)$ also scale as a function of $n$. Our goal is to establish sufficient conditions on the scaling of the triple $(n, p, d)$ such that proposed estimator is consistent:

$$P[\hat{E}_n = E^*] \to 1, \text{ as } n \to +\infty$$
Neighborhood-based logistic regression.

- Recovering $E^*$ is equivalent to recovering "signed neighborhood" for each $r \in V$

$$\mathcal{N}_\pm(r) := \{\text{sign}(\theta^*_{rt})t | t \in \mathcal{N}(r)\}$$  \hspace{1cm} (3)

where $\mathcal{N}(r)$ is the neighborhood set of $r$, i.e.,
$$\mathcal{N}(r) := \{t \in V | (r, t) \in E\}$$

- $\mathcal{N}_\pm(r)$ can be recovered from the sign-sparsity pattern:
$$\theta^*_r := \{\theta^*_{ru}, \ u \in V \setminus r\}.$$  

Derive the conditional distribution of $X_r$ given all others $X_{\setminus r}$ (4),

$$P_{\theta^*}(x_r | x_{\setminus r}) = \frac{\exp(2x_r \sum_{t \in V \setminus r} \theta^*_{rt} x_t)}{\exp(2x_r \sum_{t \in V \setminus r} \theta^*_{rt} x_t) + 1}$$  \hspace{1cm} (4)
Neighborhood-based logistic regression.

- The conditional distribution suggests that $X_r$ can be viewed as the response variable in a logistic regression in which $X_{\backslash r}$ are covariates.
- Estimation procedure: computing an $\ell_1$-regularized logistic regression of $X_r$ on $X_{\backslash r}$. Given a sample $\mathcal{X}_n^r$, the regularized regression problem is a convex program of the form

$$\min_{\theta_{\backslash r} \in \mathbb{R}^{p-1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(\theta; x^{(i)}) - \sum_{u \in V \backslash r} \theta_{ru} \hat{\mu}_{ru} + \lambda_n \| \theta_{\backslash r} \|_1 \right\}$$  \hspace{1cm} (5)

where

$$f(\theta; x) := \log \left\{ \exp \left( \sum_{t \in V \backslash r} \theta_{rt} x_t \right) + \exp \left( - \sum_{t \in V \backslash r} \theta_{rt} x_t \right) \right\}$$  \hspace{1cm} (6)
By Lagrangian duality, the problem (5) can be re-cast as a constrained problem over the ball $\|\theta_r\|_1 \leq C(\lambda n)$.

- After obtaining the minimizer $\hat{\theta}_r^n$, estimate the signed neighborhood $\mathcal{N}_\pm(r)$ according to

$$\hat{\mathcal{N}}_\pm(r) := \{\text{sign}(\hat{\theta}_{ru})u | u \in V \setminus r, \hat{\theta}_{su} \neq 0\}.$$  

- The full graph $G$ is estimated consistently $\{\hat{E}_n = E^*\}$, if $\hat{\mathcal{N}}_\pm(r) = \mathcal{N}_\pm(r)$ for all $r \in V$. 

Our analysis proceeds by

- First, establish sufficient conditions for correct signed neighborhood recovery for some fixed node $r \in V$.
- Second, we use a union bound over all $p$ nodes of the graph to conclude that consistent graph selection is also achieved.
Assumptions

Next, state the main result on the performance of neighborhood logistic regression for graphical model selection.

Definitions

- Population Fisher information matrix (Hessian of the likelihood function).

\[ Q_{r}^{*} := E_{\theta^{*}} \{ \nabla^{2} \log P_{\theta^{*}}[X_{r}|X_{\backslash r}] \}. \]

- \( S := \{(r, t)|t \in \mathcal{N}(r)\}, \)
  \( Q_{SS}^{*} \) denotes the \( d \times d \) sub-matrix of \( Q^{*} \)

- \( \theta_{\min}^{*} = \min_{(r,t) \in E} |\theta_{rt}^{*}| \) determines the limits of model selection.
Assumptions

- **(A1) Dependency condition**

\[
\lambda_{\text{min}}(Q_{SS}^*) \geq C_{\text{min}} \\
\lambda_{\text{max}}(E_{\theta}^*[X_{\backslash r} X_{\backslash r}^T]) \leq D_{\text{max}}
\]  

\( (8) \)

- **(A2) Incoherence condition**

\[
||| Q_{SS}^* Q_{SS}^{-1} |||_\infty \leq 1 - \alpha
\]

\( (9) \)
Theorem 1

Consider an Ising graphical model with parameter vector $\theta^*$ and associated edge set $E^*$ such that conditions (A1) and (A2) are satisfied by $Q^*$, and let $X_1^n$ be a set of $n$ i.i.d. samples from the model specified by $\theta^*$. Suppose that the regularization parameter $\lambda_n$ is selected to satisfy

$$\lambda_n \geq \frac{16(2 - \alpha)}{\alpha} \sqrt{\frac{\log p}{n}}$$  \hspace{1cm} (10)$$

Then there exist positive constants $L$ and $K$, independent of $(n, p, d)$, such that if

$$n > Ld^3 \log p$$  \hspace{1cm} (11)$$
Then the following properties hold with probability at least $1 - 2\exp(-k\lambda_n^2n)$

(a) For each node $r \in V$, the $\ell_1$-regularized logistic regression (5), given data $\Lambda_1^n$, has a unique solution, and so uniquely specifies a signed neighborhood $\hat{N}_\pm(r)$.

(b) For each $r \in V$, the estimated signed neighborhood $\hat{N}_\pm(r)$ correctly excludes all edges not in the true neighborhood. Moreover, it correctly includes all edges $(r, t)$ for which $|\theta^*_rt| \geq \frac{10}{C_{min}} \sqrt{d\lambda_n}$. 

Jiawei Wen
Corollary 1

Consider a sequence of Ising models satisfy conditions (A1) and (A2). For each $n$, Suppose that $(n, p(n), d(n))$ satisfies the scaling condition (11). Suppose that $\{\lambda_n\}$ of regularization parameters satisfies condition (10) and $\lambda_n^2 n \to \infty$ and the minimum parameter weights satisfy

$$\min_{(r,t) \in E^*_n} |\theta^*_{(n,p,d)}(r,t)| \geq \frac{10}{C_{\min}} \sqrt{d} \lambda n$$

(12)

for sufficiently large $n$. Then the method is model selection consistent so that if $\hat{E}_{p(n)}$ is the graph structure estimated by the method, then $P[\hat{E}_{p(n)} = E^*_p(n)] \to 1$.
The analysis required to prove Theorem 1 can be divided naturally into two parts.

- If A(1) and A(2) hold for sample Fisher information matrix

\[
Q^n := \hat{E}[\eta(X; \theta^*)X_rX_r^T] = \frac{1}{n} \sum_{i=1}^{n} \eta(x^{(i)}; \theta^*)x^{(i)}_r(x^{(i)}_r)^T \tag{13}
\]

Then the condition in Theorem 1 are sufficient to ensure the graph is recovered with high probability.

- Under the scaling condition, imposing the assumptions on \( Q^* \) guarantees(with high probability) that analogous conditions hold for \( Q^n \)
Our proof involves the explicit construction of an optimal primal-dual pair. The zero sub-gradient optimality conditions take the form

$$\nabla \ell(\hat{\theta}) + \lambda_n \hat{z} = 0 \quad (14)$$

where the dual vector $\hat{z} \in \mathbb{R}^{p-1}$ must satisfy the properties

$$\hat{z}_{rt} = \text{sign}(\hat{\theta}_{rt}) \quad \text{if} \quad \hat{\theta}_i \neq 0 \quad \text{and} \quad |\hat{z}_{rt}| \leq 1 \quad \text{otherwise}. \quad (15)$$

A pair $(\hat{\theta}, \hat{z})$ is a primal-dual optimal solution to the convex program and its dual iff (14) and (15) are satisfied.

Such an optimal primal-dual pair correctly specifies the signed neighborhood of node $r$ iff

$$\text{sign}(\hat{z}_{rt}) = \text{sign}(\theta^*_{rt}) \quad \forall (r, t) \in S := \{(r, t) \in E\} \text{ and} \quad (16)$$

$$\hat{\theta}_{ru} = 0 \quad \text{for all} \,(r, u) \in S^c := E \setminus S \quad (17)$$
Lemma 1

Shared sparsity among optimal solutions and uniqueness of the optimal solution.

**Lemma**

Suppose that there exists an optimal primal solution $\hat{\theta}$ with associated optimal dual vector $\hat{z}$ such that $\|\hat{z}_{\mathcal{S}^c}\|_\infty < 1$. Then any optimal primal solution $\tilde{\theta}$ must have $\tilde{\theta}_{\mathcal{S}^c} = 0$. Moreover, if the Hessian sub-matrix $[\nabla^2 \ell(\hat{\theta})]_{SS}$ is strictly positive definite, then $\hat{\theta}$ is the unique optimal solution.
Define the "good event": 
\[ M(\mathcal{X}_1^n) := \{ \mathcal{X}_1^n \in \{-1, +1\}^{n \times p} | Q^n \text{satisfies (A1) and (A2)} \}. \]

Proposition 1 (Fixed design).

If the event \( M(\mathcal{X}_1^n) \) holds, \( n > Ld^2 \log(p) \), and \( \lambda_n \geq \frac{16(2-\alpha)}{\alpha} \sqrt{\frac{\log p}{n}} \), then with probability at least \( 1 - 2\exp(K\lambda_n^2) \rightarrow 1 \), the following properties hold.

(a) For each node \( r \in V \), the \( \ell_1 \)-regularized logistic regression has a unique solution, and so uniquely specifies \( \hat{N}_\pm(r) \).

(b) For each \( r \in V \), the estimated signed neighborhood vector \( \hat{N}_\pm(r) \) correctly excludes all edges not in the true neighborhood. Moreover, it correctly includes all edges with \( |\theta_{rt}| \geq \frac{10}{C_{\text{in}}} \sqrt{d} \lambda_n \).
Extensions to general discrete Markov random fields.

- Each variable $X_i$ taking values in $X = \{1, 2, \cdots, m\}$
- Define $I[x_s = j] = \begin{cases} 1, & \text{if } x_s = j, \\ 0, & \text{otherwise}. \end{cases}$ Similarly, define pairwise indicator functions $I[x_s = j, x_t = k]$
- Define a matrix $\Theta^*_r \in \mathbb{R}^{(m-1)^2 \times (p-1)}$, where column $u$ is given by the vector $\theta^*_{ru}$.
- Solve the following convex program

$$
\min_{\Theta \backslash r \in \mathbb{R}^{(m-1)^2 \times (p-1)}} \left\{ \mathcal{L}(\Theta \backslash r; X_1^n) + \lambda_n \|\|\Theta^*_r\|\|_{2,1} \right\} \quad (18)
$$
Three different classes of graphs

- four-nearest neighbor lattices
- eight-nearest neighbor lattices
- star-shaped graphs

Fig. 1. Illustrations of different graph classes used in simulations. (a) Four-nearest neighbor grid (d = 4). (b) Eight-nearest neighbor grid (d = 8). (c) Star-shaped graph \( d = \Theta(p) \), or \( d = \Theta(\log(p)) \).
Generate random data sets \( \{x^{(1)}, \ldots, x^{(n)}\} \) by gibbs sampling for the lattice models, and by exact sampling for the star graph.

Examine the performance of models with mixed couplings \( \theta_{st}^* = \pm \omega \) or with positive couplings \( \theta_{st}^* = \omega \)

\[ \lambda = \sqrt{\log pn} \]

Perform simulations for sample sizes \( n = 10^\beta d \log p \), where the control parameter \( \beta \) ranged from 0.1 to 3, depending on the graph type.
Results for 4-nearest-neighbor grid model

- Three different graph size $p \in \{64, 100, 225\}$
- Each curve corresponds to the success probability versus the control parameter $\beta$.
- Each point corresponds to the average of $N = 200$ trials.

![Graphs showing success probability vs. control parameter for different graph sizes and grid types.](image)

**Fig. 2.** Plots of success probability $P[\hat{N}_\pm(r) = N(r), \forall r]$ versus the control parameter $\beta(n, p, d) = n/[10d \log(p)]$ for Ising models on 2-D grids with four nearest-neighbor interactions ($d = 4$). (a) Randomly chosen mixed sign couplings $\theta_{st}^* = \pm 0.50$. (b) All positive couplings $\theta_{st}^* = 0.50$. 

Jiawei Wen
Results for 8-nearest-neighbor grid model

**Fig. 2.** Plots of success probability $P[\hat{N}_\pm(r) = N(r), \forall r]$ versus the control parameter $\beta(n,p,d) = n/[10d \log(p)]$ for Ising models on 2-D grids with four nearest-neighbor interactions ($d = 4$). (a) Randomly chosen mixed sign couplings $\theta^*_{st} = \pm 0.50$. (b) All positive couplings $\theta^*_{st} = 0.50$. 
Results for star-shaped graphs

- Construct star-shaped graphs with $p$ vertices: one node as the hub and connecting it to $d < (p - 1)$ of its neighbors.
- For linear sparsity, we choose $d = \lceil 0.1p \rceil$; for logarithmic sparsity we choose $d = \lceil \log(p) \rceil$.
- A triple of graph sizes $p \in \{64, 100, 225\}$.

![Graph Images](image-url)
A technique based on $\ell_1$-regularized logistic regression can be used to perform consistent model selection in binary Ising graphical models.

Our analysis applies to the high-dimensional setting, in which both the number of nodes $p$ and maximum neighborhood sizes $d$ are allowed to grow as a function of the number of observations $n$.

Simulation results show the accuracy of these theoretical predictions.

The methods of this paper can be extended to general discrete graphical models with a higher number of states.
Future Work

- Our experimental results are consistent with the conjecture that logistic regression procedure fails with high probability for sample sizes $n$ that are smaller than $O(d \log p)$. It would be interesting to prove such a converse result.

- The case of samples drawn in a non-i.i.d. manner from some unknown Markov random field; we suspect that similar results would hold for weakly dependent sampling schemes.