Penalized Monte Carlo methods in high-dimensional Ising model

Wojciech Rejchel

Nicolaus Copernicus University in Toruń, Poland

Joint work with Błażej Miasojedow (University of Warsaw)

Markov random field

- Undirected graph (V, E)

Markov random field

- Undirected graph (V, E)
- $V = \{1, \ldots, d\}$ - set of vertices

Markov random field

- Undirected graph (V, E)
- $V = \{1, \ldots, d\}$ - set of vertices
- $E \subset V \times V$ - set of edges

Markov random field

- Undirected graph (V, E)
- $V = \{1, \ldots, d\}$ - set of vertices
- $E \subset V \times V$ - set of edges
- $Y = (Y(1), \ldots, Y(d))$ - random vector

Markov random field

- Undirected graph (V, E)
- $V = \{1, \ldots, d\}$ - set of vertices
- $E \subset V \times V$ - set of edges
- $Y = (Y(1), \ldots, Y(d))$ - random vector
- $Y(s)$ is associated with vertex $s \in V$

Ising model

- $Y(s) \in \{-1, 1\}$

Ising model

- $Y(s) \in \{-1, 1\}$
- Joint distribution of Y is given by

$$p(y|\theta^\star) = \frac{1}{C(\theta^\star)} \exp\left(\sum_{r<s} \theta^\star_{rs} y(r) y(s)\right)$$

Ising model

- $Y(s) \in \{-1, 1\}$
- Joint distribution of Y is given by

$$p(y|\theta^\star) = \frac{1}{C(\theta^\star)} \exp\left(\sum_{r<s} \theta^\star_{rs} y(r) y(s)\right)$$

- $\theta^\star \in \mathbb{R}^{\frac{d(d-1)}{2}}$ - true parameter

Ising model

- $Y(s) \in \{-1, 1\}$
- Joint distribution of Y is given by

$$p(y|\theta^\star) = \frac{1}{C(\theta^\star)} \exp\left(\sum_{r<s} \theta^\star_{rs} y(r) y(s)\right)$$

- $\theta^\star \in \mathbb{R}^{\frac{d(d-1)}{2}}$ - true parameter
- Intractable norming constant

$$C(\theta^\star) = \sum_{y \in \{0,1\}^d} \exp\left(\sum_{r<s} \theta^\star_{rs} y(r) y(s)\right)$$

Ising model

- $Y(s) \in \{-1, 1\}$
- Joint distribution of Y is given by

$$p(y|\theta^\star) = \frac{1}{C(\theta^\star)} \exp\left(\sum_{r<s} \theta^\star_{rs} y(r) y(s)\right)$$

- $\theta^\star \in \mathbb{R}^{\frac{d(d-1)}{2}}$ - true parameter
- Intractable norming constant

$$C(\theta^\star) = \sum_{y \in \{0,1\}^d} \exp\left(\sum_{r<s} \theta^\star_{rs} y(r) y(s)\right)$$

- $J(y) = \left(y(r)y(s)\right)_{r<s}$

Ising model

- $Y(s) \in \{-1, 1\}$
- Joint distribution of Y is given by

$$p(y|\theta^\star) = \frac{1}{C(\theta^\star)} \exp\left(\sum_{r<s} \theta^\star_{rs} y(r) y(s)\right)$$

- $\theta^\star \in \mathbb{R}^{\frac{d(d-1)}{2}}$ - true parameter
- Intractable norming constant

$$C(\theta^\star) = \sum_{y \in \{0,1\}^d} \exp\left(\sum_{r<s} \theta^\star_{rs} y(r) y(s)\right)$$

- $J(y) = \left(y(r)y(s)\right)_{r<s}$
-

$$p(y|\theta^\star) = \frac{1}{C(\theta^\star)} \exp\left[(\theta^\star)' J(y)\right]$$

Ising model

- $\theta^{\star}_{rs} = 0$

Ising model

- $\theta^{\star}_{rs} = 0$ means that $Y(r)$ and $Y(s)$ are conditionally independent

Ising model

- $\theta^\star_{rs} = 0$ means that $Y(r)$ and $Y(s)$ are conditionally independent
- Finding conditional independence

Ising model

- $\theta^{\star}_{rs} = 0$ means that $Y(r)$ and $Y(s)$ are conditionally independent
- Finding conditional independence $\Leftrightarrow$ recognizing structure of graph

Ising model

- $\theta^\star_{rs} = 0$ means that $Y(r)$ and $Y(s)$ are conditionally independent
- Finding conditional independence $\Leftrightarrow$ recognizing structure of graph $\Leftrightarrow$ estimation of $\theta^\star$

Likelihood estimation

- $Y_1, \ldots, Y_n$ - independent random vectors from $p(\cdot|\theta^\star)$

Likelihood estimation

- $Y_1, \ldots, Y_n$ - independent random vectors from $p(\cdot|\theta^\star)$
- Negative log-likelihood

$$\ell_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \theta' J(Y_i) + \log C(\theta)$$

Likelihood estimation

- $Y_1, \ldots, Y_n$ - independent random vectors from $p(\cdot|\theta^\star)$
- Negative log-likelihood

$$\ell_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \theta' J(Y_i) + \log C(\theta)$$

- Pseudolikelihood approximation

Likelihood estimation

- $Y_1, \ldots, Y_n$ - independent random vectors from $p(\cdot|\theta^\star)$
- Negative log-likelihood

$$\ell_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\theta' J(Y_i) + \log C(\theta)$$

- Pseudolikelihood approximation
- Monte Carlo (MC) approximation

Pseudolikelihood approximation

- $$p(y|\theta) = \prod_{s=1}^{d} p(y(s)|y(s-1), \ldots, y(1), \theta)$$

Pseudolikelihood approximation

- $$p(y|\theta) = \prod_{s=1}^{d} p(y(s)|y(s-1), \ldots, y(1), \theta)$$
  $$\approx \prod_{s=1}^{d} p(y(s)|y(-s), \theta)$$

Pseudolikelihood approximation

- $$p(y|\theta) = \prod_{s=1}^{d} p(y(s)|y(s-1), \ldots, y(1), \theta)$$
  $$\approx \prod_{s=1}^{d} p(y(s)|y(-s), \theta)$$
- $y(-s) = (y(1), \ldots, y(s-1), y(s+1), \ldots, y(d))$

MC approximation

- $h(y)$ - importance sampling distribution

MC approximation

- $h(y)$ - importance sampling distribution
- Norming constant

$$C(\theta) = \sum_{y \in \{0,1\}^d} \exp\left[\theta' J(y)\right]$$

MC approximation

- $h(y)$ - importance sampling distribution
- Norming constant

$$C(\theta) = \sum_{y \in \{0,1\}^d} \exp\left[\theta' J(y)\right] = \sum_{y \in \{0,1\}^d} \frac{\exp\left[\theta' J(y)\right]}{h(y)} h(y)$$

MC approximation

- $h(y)$ - importance sampling distribution
- Norming constant

$$
\begin{aligned}
C(\theta) &= \sum_{y\in\{0,1\}^d} \exp\left[\theta' J(y)\right] = \sum_{y\in\{0,1\}^d} \frac{\exp\left[\theta' J(y)\right]}{h(y)} h(y) \\
&= \mathbb{E}_{Y\sim h} \frac{\exp\left[\theta' J(Y)\right]}{h(Y)}
\end{aligned}
$$

MC approximation

- $h(y)$ - importance sampling distribution
- Norming constant

$$
\begin{aligned}
C(\theta) &= \sum_{y \in \{0,1\}^d} \exp\left[\theta' J(y)\right] = \sum_{y \in \{0,1\}^d} \frac{\exp\left[\theta' J(y)\right]}{h(y)} h(y) \\
&= \mathbb{E}_{Y \sim h} \frac{\exp\left[\theta' J(Y)\right]}{h(Y)}
\end{aligned}
$$

- Norming constant approximation

$$
\frac{1}{m} \sum_{k=1}^{m} \frac{\exp\left[\theta' J(Y^k)\right]}{h(Y^k)}
$$

MC approximation

- $h(y)$ - importance sampling distribution
- Norming constant

$$C(\theta) = \sum_{y\in\{0,1\}^d} \exp\left[\theta' J(y)\right] = \sum_{y\in\{0,1\}^d} \frac{\exp\left[\theta' J(y)\right]}{h(y)} h(y)$$
$$= \mathbb{E}_{Y\sim h} \frac{\exp\left[\theta' J(Y)\right]}{h(Y)}$$

- Norming constant approximation

$$\frac{1}{m}\sum_{k=1}^{m} \frac{\exp\left[\theta' J(Y^k)\right]}{h(Y^k)}$$

$Y^1, \dots, Y^m$ - Markov chain with stationary distribution h

MCMC approximation

- $Y^1, \ldots, Y^m$ - Markov chain with stationary distribution h

MCMC approximation

- $Y^1, \ldots, Y^m$ - Markov chain with stationary distribution h
-

$$\ell_n^m(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\theta' J(Y_i) + \log\left(\frac{1}{m}\sum_{k=1}^{m}\frac{\exp\left[\theta' J(Y^k)\right]}{h(Y^k)}\right)$$

High-dimensional setting

- $d = d_n >> n$

High-dimensional setting

- $d = d_n >> n$
- Number of parameters $= \frac{d(d-1)}{2}$

High-dimensional setting

- $d = d_n >> n$
- Number of parameters $= \frac{d(d-1)}{2}$
- Penalized empirical risk minimization

$$\ell_n^m(\theta) + \lambda|\theta|_1$$

High-dimensional setting

- $d = d_n >> n$
- Number of parameters $= \frac{d(d-1)}{2}$
- Penalized empirical risk minimization

$$\ell_n^m(\theta) + \lambda|\theta|_1$$

- $|\theta|_1 = \sum_{r<s} |\theta_{rs}|$

High-dimensional setting

- $d = d_n >> n$
- Number of parameters $= \frac{d(d-1)}{2}$
- Penalized empirical risk minimization

$$\ell_n^m(\theta) + \lambda|\theta|_1$$

- $|\theta|_1 = \sum_{r<s} |\theta_{rs}|$
- $\hat{\theta} = \arg\min_\theta \ell_n^m(\theta) + \lambda|\theta|_1$

Notations

- $\bar{d} = d(d-1)/2$

Notations

- $\bar{d} = d(d-1)/2$
- $T = \{(r,s) : \theta^{\star}_{rs} \neq 0\}$

Notations

- $\bar{d} = d(d-1)/2$
- $T = \{(r,s) : \theta^{\star}_{rs} \neq 0\}$
- $\bar{d}_0 = |T|$

Notations

- $\bar{d} = d(d-1)/2$
- $T = \{(r,s) : \theta^{\star}_{rs} \neq 0\}$
- $\bar{d}_0 = |T|$
- $Y^1, \ldots, Y^m$ - Gibbs sampler on $\{-1,1\}^d$ with stationary distribution h

Main results

Theorem

Let $\varepsilon > 0$. If

Main results

Theorem

Let $\varepsilon > 0$. If

1. cone invertibility condition is satisfied

Main results

Theorem

Let $\varepsilon > 0$. If

1. cone invertibility condition is satisfied
2. $n \geqslant C_1 \bar{d}_0^2 \log(\bar{d}/\varepsilon)$

Main results

Theorem

Let $\varepsilon > 0$. If

1. cone invertibility condition is satisfied
2. $n \geqslant C_1 \bar{d}_0^2 \log(\bar{d}/\varepsilon)$
3. $m \geqslant C_2 \frac{\bar{d}_0^2 M^2 \log(\beta_1 \bar{d}/\varepsilon)}{\beta_2}$

Main results

Theorem

Let $\varepsilon > 0$. If

1. cone invertibility condition is satisfied
2. $n \geqslant C_1 \bar{d}_0^2 \log(\bar{d}/\varepsilon)$
3. $m \geqslant C_2 \frac{\bar{d}_0^2 M^2 \log(\beta_1 \bar{d}/\varepsilon)}{\beta_2}$

then with probability at least $1 - 4\varepsilon$

$$\left|\hat{\theta} - \theta^\star\right|_\infty \leqslant C_3 \lambda,$$

Main results

Theorem

Let $\varepsilon > 0$. If

1. cone invertibility condition is satisfied
2. $n \geqslant C_1 \bar{d}_0^2 \log(\bar{d}/\varepsilon)$
3. $m \geqslant C_2 \frac{\bar{d}_0^2 M^2 \log(\beta_1 \bar{d}/\varepsilon)}{\beta_2}$

then with probability at least $1 - 4\varepsilon$

$$\left|\hat{\theta} - \theta^\star\right|_\infty \leqslant C_3 \lambda,$$

where

$$\lambda = \max\left(\sqrt{\frac{\log(\bar{d}/\varepsilon)}{n}}, M\sqrt{\frac{\log(\beta_1 \bar{d}/\varepsilon)}{\beta_2 m}}\right)$$

Main results

- $d \sim O(\exp(n^a)), \bar{d}_0 \sim O(n^b)$, if $a + 2b < 1$

Main results

- $d \sim O(\exp(n^a)), \bar{d}_0 \sim O(n^b)$, if $a + 2b < 1$

- Lasso estimator with threshold δ

$$\tilde{\theta}_{rs} = \begin{cases} \hat{\theta}_{rs} & \text{if } |\hat{\theta}_{rs}| > \delta \\ 0 & \text{if } |\hat{\theta}_{rs}| \leqslant \delta \end{cases}$$

Main results

- $d \sim O(\exp(n^a)), \bar{d}_0 \sim O(n^b)$, if $a + 2b < 1$

- Lasso estimator with threshold δ

$$\tilde{\theta}_{rs} = \begin{cases} \hat{\theta}_{rs} & \text{if } |\hat{\theta}_{rs}| > \delta \\ 0 & \text{if } |\hat{\theta}_{rs}| \leqslant \delta \end{cases}$$

- $\theta^\star_{min} = \min_{r<s} |\theta^\star_{rs}|$

Main results

Corollary

Let $\varepsilon > 0$. If conditions (1)-(3) are satisfied and $\theta^{\star}_{min}/2 \geqslant \delta \geqslant C_3\lambda$, then

$$P\left(\tilde{T} = T\right) \geqslant 1 - 4\varepsilon.$$

Related papers

- Ravikumar, P., Wainwright, M. J., Lafferty, J. - Ann. Statist. (2010)

Related papers

- Ravikumar, P., Wainwright, M. J., Lafferty, J. - Ann. Statist. (2010)
- Höfling, H., Tibshirani, R. - JMLR (2009)

Related papers

- Ravikumar, P., Wainwright, M. J., Lafferty, J. - Ann. Statist. (2010)
- Höfling, H., Tibshirani, R. - JMLR (2009)
- Guo J., Levina E., Michailidis G., Zhu J. (2010)

Related papers

- Ravikumar, P., Wainwright, M. J., Lafferty, J. - Ann. Statist. (2010)
- Höfling, H., Tibshirani, R. - JMLR (2009)
- Guo J., Levina E., Michailidis G., Zhu J. (2010)
- Jalali, A., Johnson, C. C., Ravikumar, P. K. - NIPS (2011)

Related papers

- Ravikumar, P., Wainwright, M. J., Lafferty, J. - Ann. Statist. (2010)
- Höfling, H., Tibshirani, R. - JMLR (2009)
- Guo J., Levina E., Michailidis G., Zhu J. (2010)
- Jalali, A., Johnson, C. C., Ravikumar, P. K. - NIPS (2011)
- Xue, L., Zou, H., Cai, T. - Ann. Statist. (2012)

Simulated data sets

- $d = 20, 50$

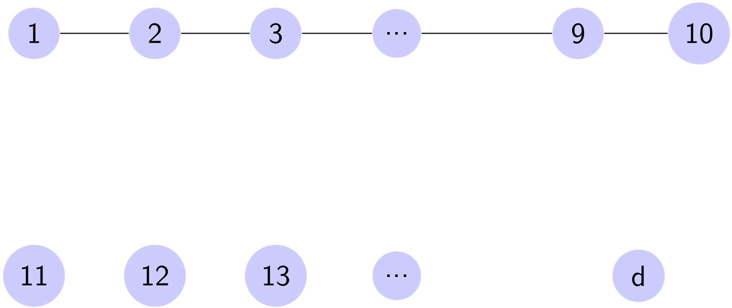
Simulated data sets

- $d = 20, 50$
- $n = 50, 100, 200, 500, 1000$

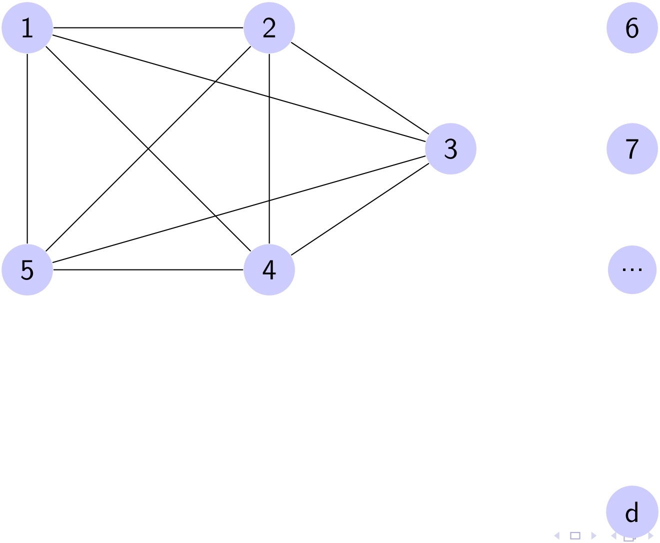
Simulated data sets

- $d = 20, 50$
- $n = 50, 100, 200, 500, 1000$
- $m = 10^5$

Model 1



Model 2



Simulated data sets

- We draw 20 configuration of signs

Simulated data sets

- We draw 20 configuration of signs
- We draw 20 replications of data set

Simulated data sets

- We draw 20 configuration of signs
- We draw 20 replications of data set
- $\lambda = c_1 * \sqrt{\log \bar{d}/n}$

Simulated data sets

- We draw 20 configuration of signs
- We draw 20 replications of data set
- $\lambda = c_1 * \sqrt{\log \bar{d}/n}$
- $\delta = c_2 * \sqrt{\log \bar{d}/n}$

Model 1

		Pseudo		MCMC	
d	n	Lasso	TL	Lasso	TL
20	50	0.23	0.37	0.02	0.18
	100	0.74	0.91	0.10	0.73
	200	0.78	1.00	0.44	0.97
	500	0.97	1.00	0.92	1.00
	1000	1.00	1.00	1.00	1.00
50	50	0.20	0.20	0.03	0.12
	100	0.70	0.83	0.07	0.61
	200	0.88	1.00	0.33	0.93
	500	0.99	1.00	0.73	1.00
	1000	1.00	1.00	0.97	1.00

Model 2

d	n	Pseudo Lasso	Pseudo TL	MCMC Lasso	MCMC TL
20	50	0.15	0.15	0.45	0.45
	100	0.14	0.14	0.51	0.51
	200	0.14	0.18	0.54	0.54
	500	0.19	0.23	0.56	0.56
	1000	0.25	0.26	0.55	0.55
50	50	0.15	0.15	0.46	0.46
	100	0.14	0.14	0.50	0.50
	200	0.15	0.15	0.53	0.53
	500	0.16	0.25	0.55	0.55
	1000	0.23	0.25	0.54	0.54

References

Besag J. (1974). *Spatial interaction and the statistical analysis of lattice systems.* J. R. Statist. Soc. B, 36, 192–236.

Guo J., Levina E., Michailidis G. and Zhu J. (2010). *Joint structure estimation for categorical Markov networks,* Technical report.

Höfling, H. and Tibshirani, R. (2009). *Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods.* J. Mach. Learn. Res., 10, 883–906.

Ising, E. (1925). *Beitrag zur theorie des ferromagnetismus.* Z. Physik, 31, 53–258.

Jalali, A., Johnson, C. C. and Ravikumar, P. K. (2011). *On learning discrete graphical models using greedy methods,* Proceedings of NIPS.

Miasojedow, B., Rejchel, W. (2016). *Sparse estimation in Ising Model via penalized Monte Carlo methods*, arXiv:1612.07497.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2010). *High-dimensional Ising model selection using* l_1*-regularized logistic regression.* Ann. Statist., 38 1287–1319.

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso.* J. R. Statist. Soc. B, 58, 267–288.

Xue, L., Zou, H. and Cai, T. (2012). *Nonconcave penalized composite conditional likelihood estimation of sparse Ising models.* Ann. Stat., 40, 1403–1429.