应用概率统计 第 35 卷
第 2 期 2019 年 4 月

Chinese Journal of Applied Probability and Statistics
Apr., 2019, Vol. 35, No. 2, pp. 165-177
doi: 10.3969/j.issn.1001-4268.2019.02.005

基于相邻选择的 Ising 模型的非凹惩罚估计 *

李凡群* 杨桂元 张孔生

(安徽财经大学统计与应用数学学院, 蚌埠, 233000)

摘 要: 本文对 Ising 模型的局部条件似然施加非凹惩罚, 得到相应参数的 Oracle 性和渐近正态性. 在一致的界下, 得到了 Ising 模型的参数矩阵的符号相合性估计, 以及在矩阵 L_1 范数下估计的收敛速度. 随机模拟和实例分析表明, 非凹惩罚估计的灵敏度普遍较高.

关键词: Ising 图模型; Logistic 回归; Lasso 估计; 非凹惩罚; 相邻选择

中图分类号: O212.4

英文引用格式: LI F Q, YANG G Y, ZHANG K S. Non-concave penalized estimation based on the neighborhood selection method for Ising model [J]. Chinese J Appl Probab Statist, 2019, 35(2): 165–177. (in Chinese)

§1. 引 言

所谓图模型是由成对集 $\mathscr{G}=\{V,E\}$ 表示, 其中 $V=\{1,2,\ldots,p\}$ 是有限顶点集, E 是存在于顶点与顶点之间的边集, 是集合 $V\times V$ 的子集. 图模型描述了 p 维随机变量 $X=(X_1,X_2,\ldots,X_p)$ 诸顶点之间的条件相依性. 高斯图模型中的参数估计与模型选择最早是由 Dempster[1] 提出的. 在离散场合下, 每一个变量 X_j 取值于相应的数集 $\mathscr{X}_j$, 则成对马尔科夫域对应的随机变量 X 有以下分布形式:

$$\mathsf{P}_\theta(x)\propto\exp\Big[\sum_{(s,t)\in E}\phi_{st}(x_s,x_t)\Big],$$

其中对于每一个边 $(s,t)\in E$, ϕ_{st} 是数对 $(x_s,x_t)\in\mathscr{X}_s\times\mathscr{X}_t$ 到实直线上的映射.

特别地, 若任意一个变量都取值于 -1 或 1, 即 $X_j\in\{-1,1\}$, 且 $\phi_{st}(x_s,x_t)=\theta_{st}x_sx_t$, 其中 θ_{st} 为一实参数, 则称之为Ising模型. 于是Ising模型中的随机变量的分布为以下形式:

$$\mathsf{P}_\theta(x)=\frac{1}{Z(\theta)}\exp\Big(\sum_{(s,t)\in E}\theta_{st}x_sx_t\Big),\tag{1}$$

其中 $Z(\theta)$ 为配分函数 (partition function), θ 是一个 $p(p-1)/2$ 维的参数, 且若 $(s,t)\in E$, 则 $\theta_{st}\neq0$, 否则 $\theta_{st}=0$.

*国家自然科学基金项目 (批准号: 11571080) 和安徽省高校自然科学项目 (批准号: KJ2017A433) 资助.

*通讯作者, E-mail: lfq006@163.com.

本文 2017 年 7 月 24 日收到, 2018 年 3 月 26 日收到修改稿.

与相关文献类似, 本文中我们采用基于 Logistic 回归的方法来分析 Ising 模型.

对 V 中的任意一个顶点 r, 用 $\mathscr{N}(r)=\{t\in V\,|\,(r,t)\in E\}$ 表示顶点 r 的相邻点集, 即对任意 $t\in\mathscr{N}(r)$, 则 $\theta_{rt}\neq 0$; 若边 (r,t) 不在边集 E 中, 则 $\theta_{rt}=0$. 另外, 用 $\theta_{\backslash r}=\{\theta_{ru}\,|\,u\in V_{\backslash r}\}$ 表示固定顶点 r 时的相应的 $p-1$ 维参数, 用 $X_{\backslash r}=\{X_t\,|\,t\in V_{\backslash r}\}$ 表示从 p 维随机变量 X 中移去变量 X_r 后余下的 $p-1$ 维随机变量.

于是在给定 $X_{\backslash r}=x_{\backslash r}$ 的条件下, X_r 的条件概率分布为

$$\mathsf{P}_\theta(x_r\,|\,x_{\backslash r})=\exp\Big(2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t\Big)\Big/\Big[1+\exp\Big(2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t\Big)\Big]. \tag{2}$$

(2) 可以用 Logistic 模型进行解释, 其中变量 X_r 是响应变量, 而其余变量 $X_{\backslash r}$ 是相应的协变量. 于是对于每个顶点 r, 可以基于模型 (2), 得到 $\theta_{\backslash r}$ 的估计, 从而恢复顶点 r 的相邻点集 $\mathscr{N}(r)$, 进而得到 Ising 模型的参数 θ 以及边集 E 的估计.

记局部条件似然为

$$L(\theta_{\backslash r};x)=\mathsf{P}_\theta(x_r\,|\,x_{\backslash r})=\exp\Big(2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t\Big)\Big/\Big[1+\exp\Big(2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t\Big)\Big],$$

则对数局部条件似然为

$$\ell(\theta_{\backslash r};x)=2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t-\ln\Big[1+\exp\Big(2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t\Big)\Big]. \tag{3}$$

于是

$$\frac{\partial\ell(\theta_{\backslash r};x)}{\partial\theta_{\backslash r}}=2x_rx_{\backslash r}\Big\{1-\exp\Big(2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t\Big)\Big/\Big[1+\exp\Big(2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t\Big)\Big]\Big\}, \tag{4}$$

并且

$$\mathsf{E}_\theta\Big[\frac{\partial\ell(\theta_{\backslash r};x)}{\partial\theta_{\backslash r}}\Big]=0. \tag{5}$$

另外,

$$\frac{\partial^2\ell(\theta_{\backslash r};x)}{\partial\theta_{\backslash r}\partial\theta_{\backslash r}^{\mathsf{T}}}=-\eta(\theta_{\backslash r};x)x_{\backslash r}x_{\backslash r}^{\mathsf{T}}, \tag{6}$$

其中

$$\begin{aligned}\eta(\theta_{\backslash r};x)&=4\exp\Big(2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t\Big)\Big/\Big[1+\exp\Big(2x_r\sum_{t\in V_{\backslash r}}\theta_{rt}x_t\Big)\Big]^2\\&=4\mathsf{P}_\theta(X_r=1\,|\,x_{\backslash r})[1-\mathsf{P}_\theta(X_r=1\,|\,x_{\backslash r})],\end{aligned}$$

所以, $\eta(\theta_{\backslash r};x)$ 恰是 X_r 的条件协方差函数, 并且 $\eta(\theta_{\backslash r};x)\leqslant 4$.

若令

$$Q_r = \mathsf{E}_\theta\Big[-\frac{\partial^2 \ell(\theta_{\backslash r}; x)}{\partial\theta_{\backslash r}\partial\theta_{\backslash r}^{\mathsf{T}}}\Big],$$

则矩阵 Q_r 是参数 $\theta_{\backslash r}$ 的基于条件分布的费歇信息阵.

用 $S = \{t\,|\,t \in \mathscr{N}(r)\}$, $S^C = \{t\,|\,t \in \mathscr{N}(r) - S\}$, 则 $Q_r(SS)$ 表示由指标集 S 所确定的 Q_r 的子矩阵, $Q_r(SS^c)$ 表示由指标集 S^C 所确定的 Q_r 的子矩阵. 在本章中, 我们用 $\Lambda_{\min}(A)$ 和 $\Lambda_{\max}(A)$ 表示矩阵 A 的最小和最大特征值. 对于 Q_r 始终有如下的假设:

A1　Q_r 是非负定的, 且 $\Lambda_{\min}(Q_r(SS)) \geqslant C_{\min}$, $\Lambda_{\max}\mathsf{E}_\theta(X_{\backslash r}X_{\backslash r}^{\mathsf{T}}) \leqslant \lambda_{\max}$, 其中 $C_{\min}$ 和 $\lambda_{\max}$ 是两个大于 0 的常数.

该假定保证与 X_r 相依的所有协变量间, 即由指标集 S 所确定的变量间不会有过分的相依性, 且 $Q_r(SS)$ 是正定的.

A2　不连贯性条件: 存在 $\alpha \in (0, 1]$, 使 $\|Q_r(S^CS)(Q_r(SS)^{-1})\|_\infty \leqslant 1 - \alpha$.

对于 $a \times b$ 的矩阵 A, $\|A\|_\infty = \max\limits_{j=1,2,\ldots,a} \sum\limits_{k=1}^{b} |A_{jk}|$. 该假设保证与 X_r 不相依的所有协变量和与 X_r 相依的所有协变量之间不存在较强的效应.

Ravikumar 等[2] 对二项数据场合下的配分函数提出对数行列式放缩, 处理 $Z(\theta)$ 的计算问题. Ravikumar 等[3] 基于 Logistic 回归, 结合 Meinshausen 和 Bühlmann[4] 提出的相邻选择方法, 在 ℓ_1 惩罚下给出了 Ising 模型的带符号的边集的 Lasso 相合估计. Barber 和 Drton[5] 基于 Chen 和 Chen[6] 提出的 EBIC 准则, 给出 Ising 图模型在 BIC 准则下的 Lasso 估计的性质. 但是基于 ℓ_1 惩罚 Lasso 估计对绝对值较大的参数有更强的惩罚, 从而导致了估计的偏差, 而 Fan 和 Li[7] 指出非凹惩罚下的变量选择具有 Oracle 性质. 本文把非凹惩罚的变量选择方法用于对 Ising 模型各顶点的相邻点集的估计, 得到相应非凹惩罚下的变量选择的性质, 并通过数值分析, 比较了相邻选择思想下的 ℓ_1 惩罚和非凹惩罚估计的结果.

本文结构如下: 在第 2 节首先基于 Ising 图模型的 Logistic 回归解释, 在非凹惩罚下, 给出了参数矩阵中各列参数的相邻选择估计相合性和 Oracle 性质. 在此基础上, 给出了 Ising 图模型的符号相合性估计, 以及在一致的界下给出了 Ising 图模型估计的收敛速度. 第 3 节利用 R 软件对三种 Ising 图模型进行随机模拟, 结果表明非凹惩罚下的相邻选择估计优于 ℓ_1 惩罚下的相邻选择估计. 第 4 节利用非凹惩罚下的相邻选择方法分析了美国中西部地区降雨量相依性图, 实例分析的结果与随机模拟的结果是吻合的. 第 5 节给出了部分引理和定理的证明.

§2. 非凹的惩罚下的 Ising 图模型的相邻选择估计

由于 Ising 图模型可由如下的对称矩阵 Θ 描述:

$$\Theta=\begin{pmatrix} 0 & \theta_{12} & \theta_{13} & \dots & \theta_{1p} \\ \theta_{21} & 0 & \theta_{23} & \dots & \theta_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_{p-1,1} & \theta_{p-1,2} & \theta_{p-1,3} & \dots & \theta_{p-1,p} \\ \theta_{p1} & \theta_{p2} & \theta_{p3} & \dots & 0 \end{pmatrix},$$

其中 $\theta_{ij}=\theta_{ji}$. 用 $\Theta_{r\backslash r}$ 表示 Θ 的第 r 列元素移除第 r 个元素后的 $p-1$ 维向量, 则 $\Theta_{r\backslash r}=\theta_{\backslash r}$.

在此记号下 (1) 可以写成

$$\mathsf{P}_\Theta(x)=\frac{1}{Z(\Theta)}\exp\Big(\frac{1}{2}x^{\mathsf{T}}\Theta x\Big).$$

2.1 Ising 图模型各列参数估计的 Oracle 性

设 $\mathfrak{X}^n=\{x^{(1)},x^{(2)},\dots,x^{(n)}\}$ 是来自与 Ising 模型 (1) 的样本容量为 n 的独立同分布的样本. 对 $\ell(\theta_{\backslash r};\mathfrak{X}^n)=\sum\limits_{i=1}^{n}\ln\mathsf{P}_\theta(x_r^{(i)}\,|\,x_{\backslash r}^{(i)})$ 的参数 $\theta_{\backslash r}$ 施行非凹的惩罚, 得到如下的 $\theta_{\backslash r}$ 的非凹的惩罚估计:

$$\widehat{\theta}_{\backslash r}=\arg\max_{\theta_{\backslash r}\in R^{p-1}}\{\ell(\theta_{\backslash r};\mathfrak{X}^n)-n\sum_{t\in V_{\backslash r}}p_\lambda(|\theta_{rt}|)\},\tag{7}$$

其中 λ 是调节参数, 且惩罚函数满足 $p_\lambda(0)=0$.

记 $Q(\theta_{\backslash r})=\ell(\theta_{\backslash r};\mathfrak{X}^n)-n\sum\limits_{t\in V_{\backslash r}}p_\lambda(|\theta_{rt}|)$, 下面的定理表明最大化 $Q(\theta_{\backslash r})$, 由 (7) 得到 $\theta_{\backslash r}$ 的相合估计.

定理 1 设 $a_n=\max\{p'_{\lambda_n}(|\theta_{rt}|):\theta_{rt}\neq0\}$, 如果 $\max\{p''_{\lambda_n}(|\theta_{rt}|):\theta_{rt}\neq0\}\to0$, Q_r 满足假设 A1, 则存在 $\widehat{\theta}_{\backslash r}$ 局部最大化 $Q(\theta_{\backslash r})$, 并且 $\|\widehat{\theta}_{\backslash r}-\theta_{\backslash r}\|_2=O_p(n^{-1/2}+a_n)$.

该定理表明, 通过选择合适的 λ_n, 存在 $\theta_{\backslash r}$ 的 $\sqrt{n}$ 相合估计.

为了进一步给出惩罚估计的 Oracle 性质, 对惩罚函数提出如下的假设:

A3 假设惩罚函数 $p_{\lambda_n}(|\theta_{\backslash r}|)$ 满足

$$\liminf_{n\to\infty}\liminf_{\theta_{\backslash r}\to0^+}p'_{\lambda_n}(\theta_{rt})/\lambda_n>0.\tag{8}$$

记 $\Sigma=\mathrm{diag}\{p''_{\lambda_n}(|\theta_{rS_1}|),p''_{\lambda_n}(|\theta_{rS_2}|),\dots,p''_{\lambda_n}(|\theta_{rS_{p-1}}|)\}$ 为对角形矩阵, 其中 S_k 表示 S 中的第 k 个指标. $b=[p'_{\lambda_n}(|\theta_{rS_1}|),p'_{\lambda_n}(|\theta_{rS_2}|),\dots,p'_{\lambda_n}(|\theta_{rS_{p-1}}|)]^{\mathsf{T}}$ 为 $p-1$ 维向量.

定理 2 (Oracle 性质) 设 $\mathfrak{X}^n=\{x^{(1)},x^{(2)},\dots,x^{(n)}\}$ 是来自于 Ising 模型 (1) 的样本容量为 n 的独立同分布的样本. 假设条件 A1 和 A3 满足, 并且如果当 $n\to\infty$ 时, $\lambda_n\to0$, $\sqrt{n}\lambda_n\to\infty$, 则 $\sqrt{n}$ 相合估计 $\widehat{\theta}_{\backslash r}=(\widehat{\theta}_{\backslash r}^{(S)},\widehat{\theta}_{\backslash r}^{(S^C)})$ 以趋向于 1 的概率满足

(i) 稀疏性: $\widehat{\theta}_{\backslash r}^{(S^C)}=0$.

(ii) 渐近正态性: $\sqrt{n}[Q_r(SS)+\Sigma]\{\widehat{\theta}^{(S)}_{\backslash r}-\theta^{(S)}_{\backslash r}+[Q_r(SS)+\Sigma]^{-1}b\}\to \mathrm{N}(0,Q_r(SS))$.

2.2　Ising 图模型估计及其收敛速度

利用第二节介绍的基于 Logistic 回归的方法, 给出 Θ 的非凹惩罚估计如下:

1. 首先利用惩罚的 Logistic 回归, 得到每个顶点的相邻集, 即由 Θ 各列的估计得到 Θ 的估计 $\widehat{\Theta}^1$.

$$\begin{aligned}\widehat{\Theta}^1_{r\backslash r}=\widehat{\theta}_{\backslash r}&=\arg\max_{\theta_{\backslash r}\in R^{p-1}}\Big\{\ell(\theta;\mathfrak{X}^n)-n\sum_{t\in V_{\backslash r}}p_{\lambda_n}(|\theta_{rt}|)\Big\}\\&=\arg\min_{\theta_{\backslash r}\in R^{p-1}}\Big\{\widetilde{\ell}(\theta;\mathfrak{X}^n)+\sum_{t\in V_{\backslash r}}p_{\lambda_n}(|\theta_{rt}|)\Big\},\end{aligned}$$

其中 $\widetilde{\ell}(\theta;\mathfrak{X}^n)=-n^{-1}\ell(\theta;\mathfrak{X}^n)$, $j=1,2,\ldots,p$.

2. 由于估计 $\widehat{\Theta}^1$ 不是对称的, 通过如下的对称化, 得到 Θ 的对称估计 $\widehat{\Theta}$.

$$\widehat{\Theta}=\arg\min_{\Theta\in\mathscr{S}_0^p}|||\Theta-\widehat{\Theta}^1|||_1, \tag{9}$$

其中 $\mathscr{S}_0^p$ 表示对角元素为0的对称矩阵族, $|||\cdot|||_1$ 表示矩阵的 L_1 范.

记与 Q_r 相应的样本费歇尔信息阵为

$$Q_r^n=\widehat{\mathsf{E}}[\eta(X;\theta)X_{\backslash r}X_{\backslash r}^{\mathsf{T}}]=\frac{1}{n}\sum_{i=1}^{n}[\eta(x^{(i)};\theta)x^{(i)}_{\backslash r}(x^{(i)}_{\backslash r})^{\mathsf{T}}].$$

带符号的图模型的边集为

$$\boldsymbol{E}=\begin{cases}\mathrm{sign}(\theta_{st}), & (s,t)\in E;\\ 0, & \text{其他}.\end{cases}$$

相应的图模型的带符号边集估计为

$$\widehat{\boldsymbol{E}}_n=\begin{cases}\mathrm{sign}(\widehat{\theta}_{st}), & (s,t)\in E;\\ 0, & \text{其他}.\end{cases}$$

若记 $W_r^n=-\partial\widetilde{\ell}(\theta;\mathfrak{X}^n)/\partial\theta_{\backslash r}$, 则我们有如下的引理:

引理 3　对于指定的常数 K, $a_n=\max\{p'_{\lambda_n}(|\theta_{rt}|):\theta_{rt}\neq 0\}$, 对任意顶点 r 有

$$\mathsf{P}(|W_r^n|_\infty\geqslant Ka_n)\leqslant 2\exp\Big[-\frac{K^2a_n^2n}{8}+\ln(p)\Big],$$

并且当 $a_n>(4/K)\sqrt{\ln(p)/n}$ 时, 该概率以速度 $\exp\{-ca_n^2n\}$ 收敛到 0.

引理 4 记 $Q_\alpha = \nabla^2\widetilde{\ell}((\theta_S + \alpha u_S); \mathfrak{X}^n)$, $\alpha \in [0,1]$, $Q_r(SS)$ 满足条件 A1, 且对指定的常数 M, $\|u_S\|_2 = M\lambda_n\sqrt{d}$, 则

$$\Lambda_{\min}(Q_\alpha) \geqslant \frac{1}{2}C_{\min}.$$

该引理的证明来自于文献 [3] 的引理 3 的证明过程.

于是当样本费歇信息阵满足一定的条件时, Ising 图模型的估计 $\widehat{\Theta}$ 有如下性质.

定理 5 设 $\mathfrak{X}^n = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ 为来自 Ising 模型的样本, 对任意顶点 r, Q_r^n 满足假设 A1, $b_n = \max\{|p'_{\lambda_n}(|\theta_{rt}|)|/\lambda_n : \theta_{rt} \neq 0\}$, $c_n = \max\{|p''_{\lambda_n}(|\theta_{rt}|)| : \theta_{rt} \neq 0\}$, $\theta_{\min} = \min\{|\theta_{st}|, (s,t) \in E\} > \lambda_n\sqrt{d}(4b_n+1)/C_{\min}$, $C_{\min} \geqslant 2c_n$, 则非凹惩罚估计 $\widehat{\Theta}$ 以概率不小于 $1 - \exp\{-K^2a_n^2n/8 + \ln(p)\}$ 有

(i) 符号相合性: $\widehat{\boldsymbol{E}}_n = \boldsymbol{E}$.

(ii) 误差收敛速度: $|||\widehat{\Theta} - \Theta|||_1 \leqslant \lambda_n d(4b_n+1)/C_{\min}$.

§3. 随机模拟

本节中, 我们考察非凹惩罚下的 Ising 模型估计的表现, 考虑如下三种模型:

• 模型 1 随机模型 (Random graph): 任意两顶点间以 0.3 的概率存在边, 以 0.7 的概率不存在边来构造模型结构. 然后随机产生图模型中反映边 (i,j) 强弱的参数 θ_{ij}, 得到加权 (Weighted) 的图模型.

• 模型 2 4 相邻模型 (nearest edges graph): 该模型是把 p 个顶点安排在 $\sqrt{p} \times \sqrt{p}$ 的网格内, 每个顶点与其上下左右四个顶点关联, 即每个顶点与其上下左右的四个顶点之间存在 4 条边, 该模型的最大模型度 (degree) 为 4. 在我们的模拟研究中取边 $E(i,j)$ 对应的参数 $\theta_{ij} = 0.5$.

• 模型 3 星形模型 (Star graph): 模型中的边由某个顶点与其余的 $p-1$ 个顶点中 q 个的顶点组成, $q \leqslant p-1$. q 用两种方式确定: (i) $q = [\ln(p)]$ (对数稀疏图); (ii) $q = [ap]$ (线性稀疏图). 在后面的模拟中, 我们仅仅报告了线性稀疏图的模拟结果, 其中 $a = 0.3$, 并且取边 $E(i,j)$ 对应的参数 $\theta_{ij} = 0.75$.

Ravikumar 等[3] 用 ℓ_1 惩罚的方法研究了模型 2 和模型 3 的图的估计, Barber 和 Drton[5] 基于 ℓ_1 惩罚, 讨论了 BIC 准则下估计的性质, 并且也对模型 2 和模型 3 进行了随机模拟研究. 我们利用 R 软件进行编程求解: 首先利用 R 的 IsingSampler 程序包产生数据, 然后借助 R 的 IsingFit 和 ncvreg 程序包实现估计.

在实际模拟过程中, 我们取 SCAD 为惩罚函数, 分别考虑 BIC 和 CV 准则选取调节参数, 从而进行模型选择. 其中

• BIC 准则

记

$$\begin{aligned}\mathrm{BIC}_\lambda &= -2\ln L(\widehat{\theta}_{\backslash r}) + d_r\ln(n) + 2\lambda\ln(p)\\ &= -2\sum_{i=1}^{n}\{x_r^{(i)}(x_{\backslash r}^{(i)})^{\intercal}\widehat{\theta}_{\backslash r} - b[(x_{\backslash j}^{(i)})^{\intercal}\widehat{\theta}_{\backslash r}]\} + d_r\ln(n) + 2\lambda\ln(p),\end{aligned}$$

其中 d_r 为估计 $\widehat{\theta}_{\backslash r}$ 中非零元的个数, $b[(x_{\backslash r}^{(i)})^{\intercal}\widehat{\theta}_{\backslash r}] = \ln(1+\mathrm{e}^{x_r^{(i)}(x_{\backslash r}^{(i)})^{\intercal}\widehat{\theta}_{\backslash r}})$.

根据 (2), 对每一个 r, 考虑 X_r 对其余变量 $X_{\backslash r}$ 的 Logidtic 回归模型, 利用 BIC 准则得到 X_r 对其余变量 $X_{\backslash r}$ 的 Logidtic 回归系数 $\widehat{\theta}_{\backslash r}$ 为: $\widehat{\theta}_{\backslash r} = \arg\min_{\theta_{\backslash r}\in R^{p-1}} \mathrm{BIC}_\lambda$.

• CV 准则

首先把样本数据 T 随机分成 K 组, T^k 表示第 k 组数据, 把它作为检验数据, 其余数据 $T - T^k$ 作为训练样本, 由训练样本数据得到惩罚估计 $\widehat{\theta}^k_{\backslash r}$, 定义 CV 函数如下:

$$\mathrm{CV}(\lambda) = \sum_{k=1}^{K}\sum_{(x^{(i)})\in T^k}\{x_r^{(i)}(x_{\backslash r}^{(i)})^{\intercal}\widehat{\theta}^k_{\backslash r} - b[(x_{\backslash r}^{(i)})^{\intercal}\widehat{\theta}^k_{\backslash r}]\},$$

则在 CV 准则下的 λ 为 $\lambda(\mathrm{CV}) = \arg\max_{\lambda}\mathrm{CV}(\lambda)$. 最后, 取 $\lambda = \lambda(\mathrm{CV})$, 利用全体数据得到 $\theta_{\backslash r}$ 的估计.

我们把两种准则下非凹惩罚的图模型的估计结果同基于 ℓ_1 惩罚的 BIC 估计结果进行比较. 基于 ℓ_1 惩罚的估计是通过 IsingFit R package 实现, 非凹惩罚的图模型的估计是通过 ncvreg R package 来实现. 由于 ncvreg 是针对于 lodistic 回归问题的, 而在 Ising 模型场合下, 由 (2) 式, 我们在用该程序包时对数据进行如下的变换: 当考虑 X_j 对其余变量 $X_{\backslash j}$ 的 logidtic 回归时, 把响应变量记为 $Y_j = (X_j+1)/2$, 由此得到的估计记为 $\widehat{\beta}$, 最后, 得到参数估计为 $\widehat{\theta}_{\backslash r} = \widehat{\beta}/2$.

在模拟中, 我们用通过相邻选择获得诸 $\theta_{\backslash r}$ 的估计, 然后分别用 “AND” 和 “OR” 准则, 直接给出了 Θ 的估计. 在模拟的结果中, 我们比较了不同方法下的估计的 “Sensitivity (灵敏度)” 和 “Specificity (特异度)” [2]. 灵敏度指的是真实的模型中存在边被估计为边的比例; 特异度是指真实的图模型中不存在的边被估计为不存在的比例. 由于调节参数 $\lambda \propto \gamma\sqrt{\ln(p)/n}$, 其中 γ 称为控制参数, 则样本量 $n = \gamma\ln(p)/\lambda^2 \triangleq 10\gamma\ln(p)$. 图 1 – 图 6 分别给出了 “灵敏度” 和 “特异度” 对控制参数 γ 的曲线图. 其中曲线 “SEN-CV-AND-rule” 和 “SEN-BIC-AND-rule” 以及 “SPE-CV-AND-rule” 和 “SPE-BIC-AND-rule” 是由本文提出的非凹惩罚下的估计结果; “SEN-L1-AND-rule” 和 “SPE-L1-AND-rule” 是由 ℓ_1 惩罚下的 Lasso 估计的结果.

从图中可以看到, 基于非凹惩罚下的惩罚估计在 CV 准则下的估计的灵敏度方面一致比基于 ℓ_1 惩罚的 Lasso 估计要好, 在模型的估计的特异度方面, 两者表现相当. 基于非凹惩罚下的惩罚估计在 BIC 准则下的估计的灵敏度方面一致比基于 ℓ_1 惩罚的 Lasso 估计要好, 提高了估计模型的特异度, 但是牺牲了部分估计的灵敏度.

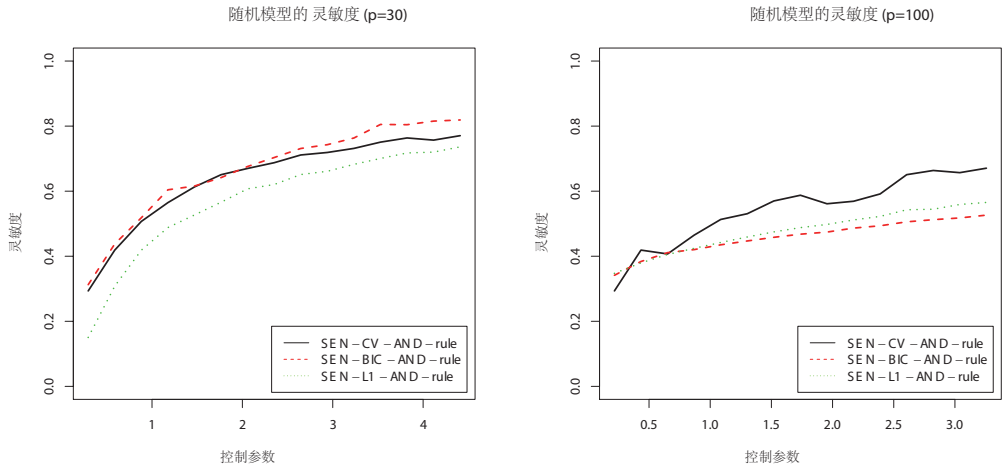

图 1　从左到右分别为模型 1 中灵敏度对控制参数的曲线图

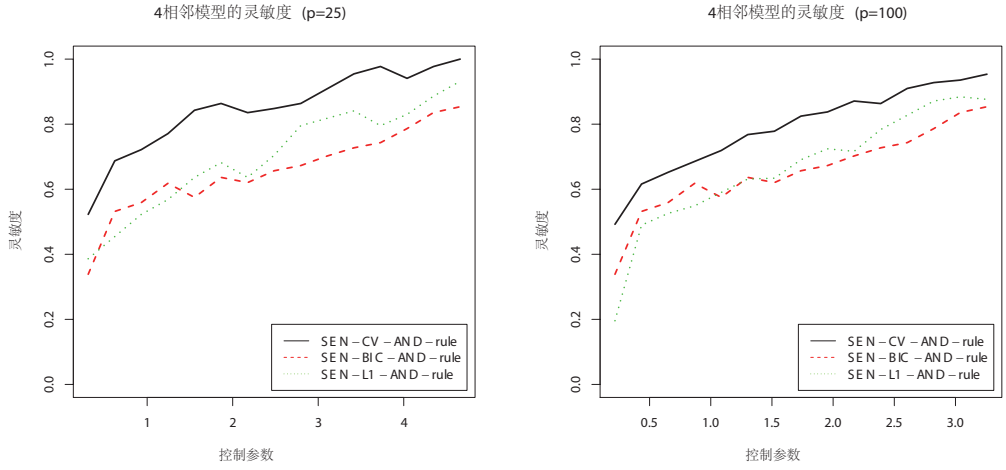

图 2　从左到右分别为模型 2 中灵敏度对控制参数的曲线图

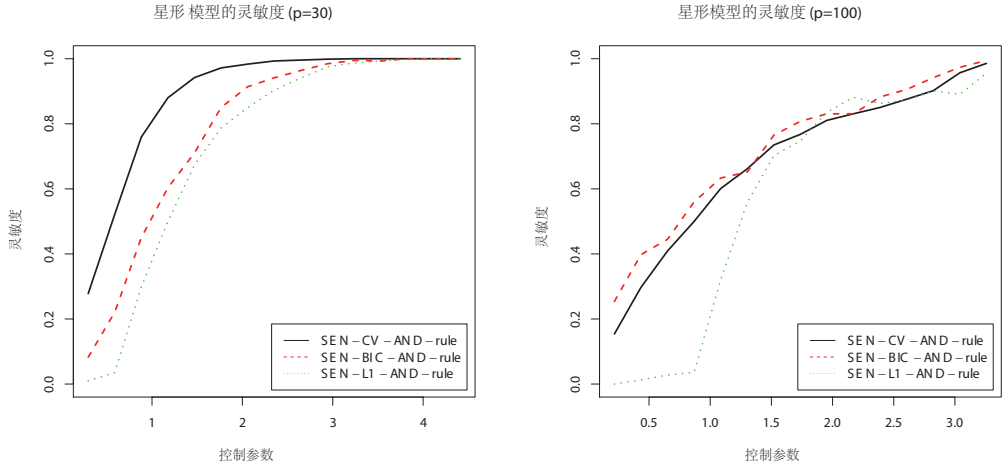

图 3　从左到右分别为模型 3 中灵敏度对控制参数的曲线图

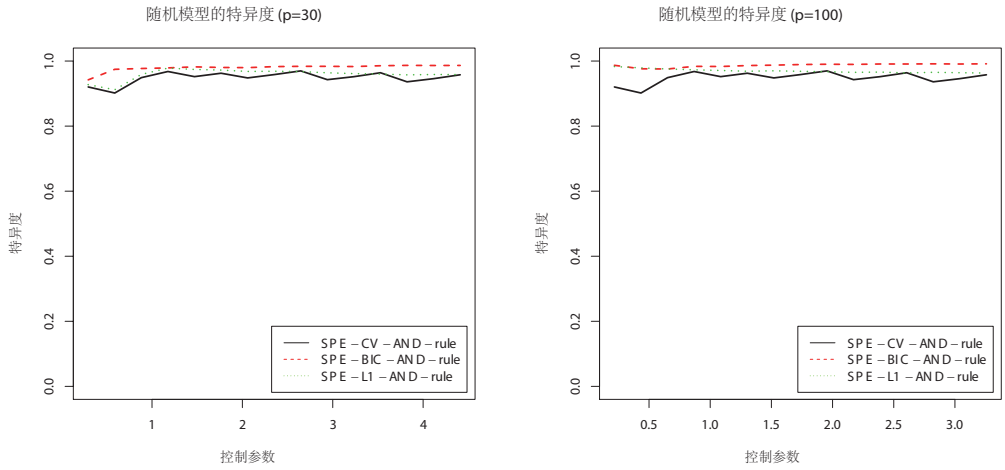

图 4　从左到右分别为模型 1 中特异度对控制参数的曲线图

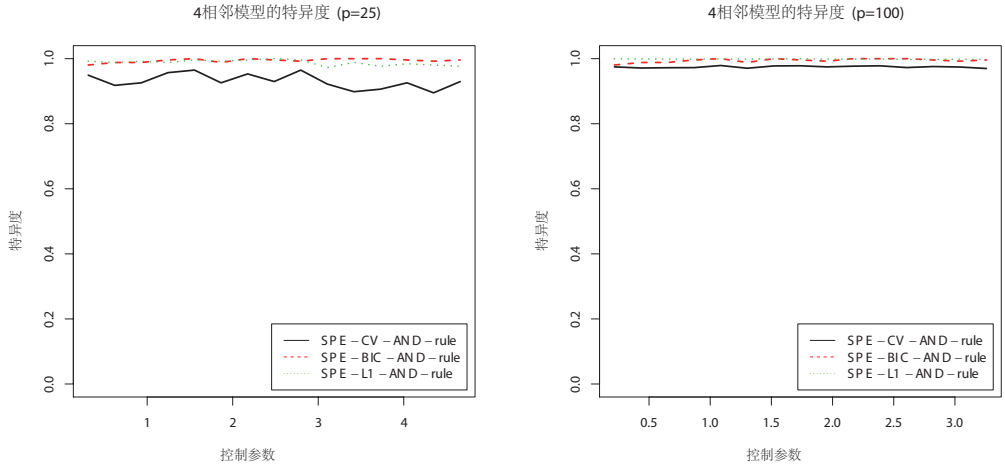

图 5　从左到右分别为模型 2 中特异度对控制参数的曲线图

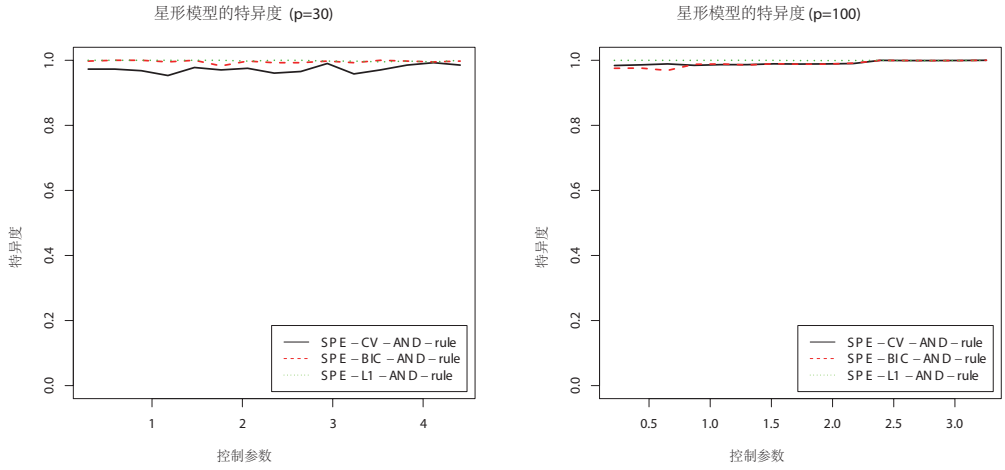

图 6　从左到右分别为模型 3 中特异度对控制参数的曲线图

§4. 实例分析

在该例中, 我们利用 Ising 模型, 分析美国中西部地区的降雨情况的相依性. 用二项数据描述降雨情况: 若某天降雨, 则记为 1, 否则记为 0. 实际数据来自于网站 “The United States Historical Climatology Network” (Menne, Willianms Jr. 和 Vose, 2011). 该数据给出了 1890 年到 2014 年每日的天气情况, 包括: 每日的最高气温、最低气温、降雪量以及降雨量. Barber 和 Drton[5] 利用基于 ℓ_1 惩罚的 Ising 模型分析了中西部四个州的部分地区的降雨情况的相依性. 这四个州分别是: Illinois, Indiana, Iowa, Missouri.

考虑到数据的完整性和连续性, 我们选取该四个州的 54 个地区 1994 年 – 2014 年间的降雨量数据, 从每个月选取两天的降雨量作为变量, 这两天之间的间隔取得较大, 以便尽可能地消除同一地区变量间 (降雨) 的相依性. 最后我们得到了每个地区的 502 天的降雨量数据, 即得到了维度 $p = 54$, 样本量为 $n = 502$ 的样本. 基于该选取的数据集, 分别用 ℓ_1 惩罚的 Ising 模型和非凹惩罚的 Ising 模型, 分析该 54 个地区的降雨相依性网络. 由于缺少真实的网络图作为参照, 和 Barber 和 Drton[5] 一样, 我们把由 54 个地区的经纬度确定的 Delaunay 三角剖分网作为真实的他们降雨相依性网.

我们先由 54 个地区的经纬度, 做出它们的 Delaunay 三角剖分网, 再由该剖分网给出相应的地区之间的降雨情况的无向图, 结果由图 7 给出. 然后分别计算了 EBIC 准则下的 Ising 模型的 ℓ_1 惩罚估计和 BIC 准则以及 CV 准则下的非凹惩罚估计的灵敏度 (正确识别 Delaunay 三角剖分网中的存在的边的比), 并以此比较不同方法下正确识别边的能力. 以取每月的第二天和第 16 天的数据为例, 上述三者的灵敏度分别为: 0.580, 0.651, 0.620. 从该结果表明, CV 准则下的非凹惩罚估计在选择正确的边的方面仍有优势, 这和前面的模拟结果也是吻合的. 图 8 给出了在 “AND-rule” 场合下的 Ising 图模型的 ℓ_1 惩罚估计和非凹惩罚估计.

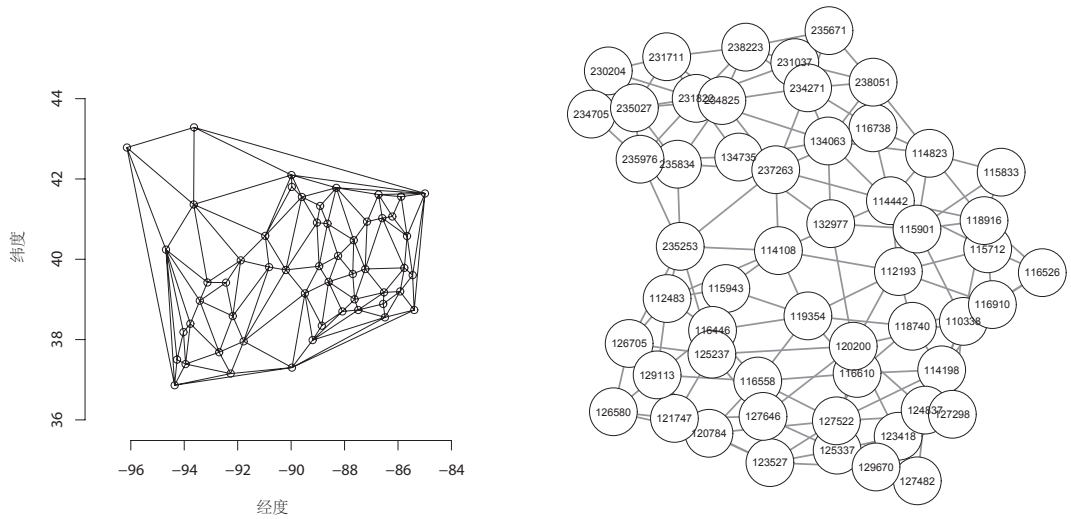

图 7　左图为 Delaunay 三角剖分网, 右图为 54 个地区基于该三角剖分网的无向图

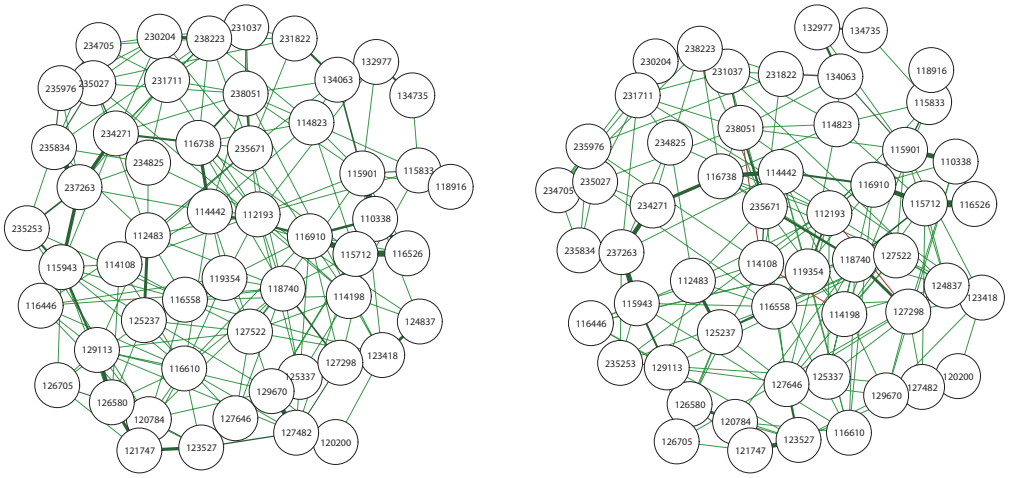

图 8　左图为 ℓ_1 惩罚下的网络估计图, 右图为非凹惩罚下的网络估计图

§5. 证　　明

定理 1、2 的证明:　定理的证明可以参照文献 [7] 非凹惩罚下的变量选择 Oracle 性的证明, 此处略去. □

引理 3 的证明:　由于 $W_r^n = -\partial\widetilde{\ell}(\theta;\mathfrak{X}^n)/\partial\theta_{\backslash r}$, 则 $W_{r,u}^n = n^{-1}\sum\limits_{i=1}^{n} Z_u^{(i)}$, 其中 $Z_u^{(i)} = x_{\backslash r,u}^{(i)}[x_r^{(i)} - \mathsf{P}_\theta(X_r^{(i)} = 1 \,|\, x_{\backslash r}^{(i)}) + \mathsf{P}_\theta(X_r^{(i)} = -1 \,|\, x_{\backslash r}^{(i)})]$.

而

$$\begin{aligned}\mathsf{E}(Z_u^{(i)}) &= \mathsf{E}[\mathsf{E}(Z_u^{(i)} \,|\, x_{\backslash r}^{(i)})]\\ &= \mathsf{E}[x_{\backslash r,u}^{(i)}\mathsf{E}(x_r^{(i)} \,|\, x_{\backslash r}^{(i)}) - \mathsf{P}_\theta(X_r^{(i)} = 1 \,|\, x_{\backslash r}^{(i)}) + \mathsf{P}_\theta(X_r^{(i)} = -1 \,|\, x_{\backslash r}^{(i)})]\\ &= 0.\end{aligned}$$

又由于 $|Z_u^{(i)}| \leqslant 2$, 所以 $Z_u^{(i)}$ $(i = 1, 2, \ldots, n)$ 是独立同分布的, 均值为零的有界随机变量. 由 Azuma-Hoeffding 不等式, 对 $\forall\, \delta > 0$, 有

$$\mathsf{P}(|W_{r,u}^n| > \delta) \leqslant 2\exp\Big(-\frac{n\delta^2}{8}\Big),$$

令$\delta = ka_n$, 则

$$\mathsf{P}(|W_r^n|_\infty > \delta) \leqslant 2\exp\Big(-\frac{nK^2a_n^2}{8}\Big) = 2p\exp\Big[-\frac{nK^2a_n^2}{8} + \ln(p)\Big],$$

引理得证. □

定理 5 的证明:　为了表述上的方便, 对任意顶点 r, 我们记 $\theta_S = \theta_{\backslash r}^{(S)}$.

对任意顶点 r,

$$G(u_S)=\widetilde{\ell}(\theta_S+u_S;\mathfrak{X}^n)-\widetilde{\ell}(\theta_S;\mathfrak{X}^n)+p_{\lambda_n}(|\theta_S+u_S|)-p_{\lambda_n}(|\theta_S|),$$

根据定理 1 和定理 2 的稀疏性知, 由 (7) 得到的 $\widehat{\theta}_S$ 使 $\widehat{u}_S=\widehat{\theta}_S-\theta_S$ 最小化 $G(u_S)$, 并且 $G(\widehat{u}_S)<0$.

若对某些常数 M, 当 $\|u_S\|_2=M\lambda_n\sqrt{d}$ 时, 有 $G(u_S)>0$, 则由 $G(u_S)$ 的凸性, 必有 $\|\widehat{u}_S\|_2<M\lambda_n\sqrt{d}$. 下面求满足条件 $G(u_S)>0$ 的 M 的界.

对 $G(u_S)$ 进行泰勒展开, 有

$$G(u_S)=(W_S^n)^{\mathrm{T}}u_S+u_S^{\mathrm{T}}\widetilde{\ell}''(\theta_S+\alpha u_S)u_S+[p'_{\lambda_n}(|\theta_S|)]^{\mathrm{T}}u_S+u_S^{\mathrm{T}}p''_{\lambda_n}(|\theta_S+\alpha u_S|)u_S.$$

由引理 3 和引理 4 有

$$|W_S^n|\leqslant\|u_S\|_\infty\|u_S\|_1\leqslant\frac{\lambda_n^2dM}{4},$$

$$u_S^{\mathrm{T}}\widetilde{\ell}''(\theta_S+\alpha u_S)u_S>\frac{1}{2}C_{\min}\|u_S\|_2^2<\frac{1}{2}C_{\min}\lambda_n^2dM^2,$$

$$|[p'_{\lambda_n}(|\theta_S|)]^{\mathrm{T}}u_S|\leqslant\|p'_{\lambda_n}(|\theta_S|)\|_\infty\|u_S\|_1\leqslant b_n\lambda_n^2dM,$$

$$|u_S^{\mathrm{T}}p''_{\lambda_n}(|\theta_S+\alpha u_S|)u_S|\leqslant c_n\|u_S\|_2^2\leqslant c_n\lambda_n^2dM,$$

从而有

$$G(u_s)\geqslant-\frac{\lambda_n^2dM}{4}-b_n\lambda_n^2dM-c_n\lambda_n^2dM^2+\frac{1}{2}C_{\min}\lambda_n^2dM^2.$$

所以取

$$M>\frac{1/4+b_n}{C_{\min}/2-c_n}>\frac{4b_n+1}{2C_{\min}},$$

有

$$\|\widehat{\theta}_S-\theta_S\|_2=\|u_S\|_2<\frac{4b_n+1}{2C_{\min}}\lambda_n\sqrt{d}.$$

于是当 $\theta_{\min}=\min\{|\theta_{st}|,(s,t)\in E\}>(4b_n+1)\lambda_n\sqrt{d}/C_{\min}$ 时, 有

$$\|\widehat{\theta}_S-\theta_S\|_\infty\leqslant\|\widehat{\theta}_S-\theta_S\|_2\leqslant\frac{1}{2}\theta_{\min},$$

所以对任意边 $(s,t)\in E$, 都有 $\mathrm{sign}(\widehat{\theta}_{st})=\mathrm{sign}(\theta_{st})$, 从而 $\widehat{\boldsymbol{E}}_n=\boldsymbol{E}$.

对于 (ii), 由于 $|||\widehat{\Theta}-\Theta|||_1=|||\widehat{\Theta}-\widehat{\Theta}^1+\widehat{\Theta}^1-\Theta|||_1$, 由三角不等式知

$$|||\widehat{\Theta}-\Theta|||_1\leqslant|||\widehat{\Theta}-\widehat{\Theta}^1|||_1+|||\widehat{\Theta}^1-\Theta|||_1.$$

由 (9) 式的定义知

$$|||\widehat{\Theta}-\widehat{\Theta}^1|||_1\leqslant|||\widehat{\Theta}^1-\Theta|||_1,$$

故对于某统一的调节参数 λ_n,

$$\begin{aligned}|||\widehat{\Theta}-\Theta|||_1 &\leqslant 2|||\widehat{\Theta}^1-\Theta|||_1 \leqslant 2\sqrt{d}\max\{\|\widehat{\theta}_{\backslash r}-\theta_{\backslash r}\|_2,\, r=1,2,\ldots,p\}\\ &\leqslant 2\sqrt{d}\frac{4b_n+1}{2C_{\min}}\lambda_n\sqrt{d}=\frac{4b_n+1}{C_{\min}}\lambda_n d. \qquad \square\end{aligned}$$

参 考 文 献

[1] DEMPSTER A P. Covariance selection [J]. *Biometrics*, 1972, **28(1)**: 157–175.

[2] RAVIKUMAR P, WAINWRIGHT M J, RASKUTTI G, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence [J]. *Electron J Stat*, 2011, **5**: 935–980.

[3] RAVIKUMAR P, WAINWRIGHT M J, LAFFERTY J D. High-dimensional Ising model selection using ℓ_1-regularized logistic regression [J]. *Ann Statist*, 2010, **38(3)**: 1287–1319.

[4] MEINSHAUSEN N, BÜHLMANN P. High-dimensional graphs and variable selection with the Lasso [J]. *Ann Statist*, 2006, **34(3)**: 1436–1462.

[5] BARBER R F, DRTON M. High-dimensional Ising model selection with Bayesian information criteria [J]. *Electron J Stat*, 2015, **9(1)**: 567–607.

[6] CHEN J H, CHEN Z H. Extended Bayesian information criteria for model selection with large model spaces [J]. *Biometrika*, 2008, **95(3)**: 759–771.

[7] FAN J Q, LI R Z. Variable selection via nonconcave penalized likelihood and its oracle properties [J]. *J Amer Statist Assoc*, 2001, **96(456)**: 1348–1360.

Non-Concave Penalized Estimation Based on the Neighborhood Selection Method for Ising Model

LI Fanqun　　YANG Guiyuan　　ZHANG Kongsheng

(*College of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, 233000, China*)

Abstract: In this paper, we put non-concave penalty on the local conditional likelihood. We obtain the oracle property and asymptotic normal distribution property of the parameters in Ising model. With a union band, we obtain the sign consistence for the estimator of parameter matrix, and the convergence speed under the matrix L_1 norm. The results of the simulation studies and a real data analysis show that the non-concave penalized estimator has larger sensitivity.

Keywords: Ising graphical model; Logistic regression; Lasso estimation; nonconcave penalty; neighborhood selection

2010 Mathematics Subject Classification: 62F12