广义线性模型组 LASSO 路径算法

马景义*, 张辛连, 苏治, 刘怡文

中央财经大学统计与数学学院, 北京 100081

E-mail: jingyima@163.com, coral90@163.com, suzhi1218@163.com, finger_seven@hotmail.com

收稿日期: 2013-06-23; 接受日期: 2014-06-26; * 通信作者

国家自然科学基金 (批准号: 71403310)、北京高等学校青年英才计划、中央财经大学青年科研创新团队、中央财经大学学科建设基金资助项目

摘要 广义线性模型组 LASSO (least absolute shrinkage and selection operator) 路径 $\hat{\beta}(\lambda)$ 的计算有两项核心内容: 选择路径参数 λ 的取值; 计算组 LASSO 估计, 即给定 λ 值的 $\hat{\beta}(\lambda)$. 目前, 在广义线性模型组 LASSO 路径的计算中, 使用格点法选择 λ 值, 基于广义线性模型似然函数一阶 Taylor 近似的坐标下降算法则常用于计算组 LASSO 估计. 本文给出的广义线性模型组 LASSO 路径算法由两个子算法组成: 第一个子算法的目的是选出使得活跃集恰好改变的 λ 值; 第二个子算法是计算组 LASSO 估计的二阶近似坐标下降算法. 模拟和实际数据分析均表明, 第一个子算法能高效地发现使得活跃集恰好改变的 λ 值, 相比基于广义线性模型似然函数一阶 Taylor 近似的坐标下降算法, 本文的二阶近似算法有较明显的速度优势.

关键词 组 LASSO 广义线性模型 正则化路径 坐标下降 二阶近似

MSC (2010) 主题分类 62J07

1 引言

在有监督学习中, 预测精度通常是模型优劣的评判准则, 基于此, 产生了大量的现代建模算法, 以岭回归为代表的正则化方法, 试图通过收缩回归系数、平衡模型预测方差的减小和预测偏差的增加, 进而提高最终模型的预测精度; 在一些应用问题里, 揭示自变量集 $\mathbf{x}$ 对因变量 y 的解释作用也是建模的重点, 此时, “节俭”就成为模型选择的另一重要依据. Tibshirani^[1] 的 LASSO 方法除了能够收缩模型系数, 还具有变量选择的作用, 是最近 20 年里统计学习领域的研究热点之一. 普通回归中, Yuan 和 Lin^[2] 把 LASSO 推广为组 LASSO (group LASSO), 在变量选择时, 能够保持变量的组结构; Meier 等人^[3] 将组 LASSO 拓展至 logistic 回归. 考虑到变量个数 p 大于样本量 n 时 LASSO 最多能选出 n 个变量, 以及自变量存在组效应 (group effect) 时不能正确选择变量等问题, Zou 和 Hastie^[4] 给出了综合 LASSO 和岭回归优势的 elastic net 方法; 在大样本下, LASSO 并不满足 oracle 性质, 即变量选择的相合性和模型系数估计的一致性, Zou^[5] 提出了满足 oracle 性质的自适应 LASSO; 进一步, Wang 和 Leng^[6] 将自适应 LASSO 推广到了具有组结构的场合, 并证明了自适应组 LASSO 也具有 oracle 性质.

上述方法中, 首要的工作是计算正则化路径 (regularization path)

$$\hat{\beta}(\lambda) = \arg \min_{\beta} \Gamma(\beta; \lambda) = \arg \min_{\beta} (L(\beta) + \lambda T(\beta)), \quad (1.1)$$

英文引用格式: Ma J Y, Zhang X L, Su Z, et al. An algorithm for the estimation of regularization paths of generalized linear models with group LASSO penalty (in Chinese). Sci Sin Math, 2015, 45: 1725–1738, doi: 10.1360/N012013-00101

其中 $L(\beta)$ 为损失函数, 衡量在样本上, 自变量集 x 的模型对因变量 y 的拟合效果, λ 为路径参数, $T(\beta)$ 为罚函数, 不同的罚函数选择对应不同类型的收缩估计方法; 其次, 通过交叉验证等工具进行模型选择, 或者变量选择.

在普通回归问题中, LASSO 正则化路径的计算相对容易, Osborne 等人^[7] 证明了 LASSO 的系数估计路径具有分段线性 (piece-wise linear) 性质, 意味着只需要找到路径变点的位置, 就可以给出精确的 $\hat{\beta}(\lambda)$. 进一步, 把 LASSO 作为一个特例, Efron 等人^[8] 给出了更简洁的最小角度回归 (least angle regression, LAR). 尽管广义线性模型的 LASSO 路径不具有分段线性性质, Park 和 Hastie^[9] 仍然将最小角度回归的思路推广到计算广义线性模型的 LASSO 路径中, 寻找那些 λ 值, 恰好有变量进入或者退出模型, 找到这些 λ 值后, 计算相应的 LASSO 估计, 然后, 线性近似整个路径 $\hat{\beta}(\lambda)$. 在广义线性模型组 LASSO 路径计算中, 普遍使用格点法^[10] 选择 λ 值; 如果能推广 Park 和 Hastie^[9] 的思路, 用较少的 λ 值, 把自变量进入或者退出模型的顺序有效地刻画出来, 就可以精确地选择变量. 在高维问题中, LASSO 类方法的一个重要作用即变量选择^[11], 上述性质尤为重要.

LASSO 或者组 LASSO 估计, 即给定 λ 后, 在 $\hat{\beta}(\lambda)$ 的计算中, 一般采用 Tseng 和 Yun^[12] 的坐标下降 (coordinate descent) 方法. 在普通线性回归问题中, Fu^[13] 的 shooting 算法, Yuan 和 Lin^[2] 的组 LASSO 估计算法均可视为坐标下降方法的特例. Meier 等人^[3] 把坐标下降算法运用到 logistic 回归的组 LASSO 估计的计算中, 在每一步迭代中使用 $L(\beta)$ 的一阶导给出下降方向, 然后按照 Armijo 规则选出步长. 该做法仅利用了 $L(\beta)$ 的一阶导信息, 可能会使得使用 Armijo 规则选择步长的计算步骤较多, 步长较小, 目标函数的下降较慢. 在极端情形下, 由于步长太小, 以至于在计算机中被表达为 0, 使得算法不能按理论预期收敛, 进一步的讨论见第 2.3 小节.

本文的工作有两个重点. 首先, 推广 Park 和 Hastie^[9] 选择 λ 值的算法到广义线性模型组 LASSO 路径的计算中, 并利用模拟数据和实际数据, 分析其相比格点法的优势; 其次, 给出基于 $L(\beta)$ 一阶导和二阶导信息的坐标下降算法, 计算给定 λ 值的组 LASSO 估计 (为了区别, 本文中把利用 $L(\beta)$ 一阶导计算下降方向的坐标下降算法称为一阶近似坐标下降算法, 本文的算法为二阶近似坐标下降算法). 本文给出的广义线性模型组 LASSO 算法可以很容易地推广至含组结构约束的 elastic net 和自适应组 LASSO 等正则化方法中.

本文结构如下: 第 2 节给出广义线性模型组 LASSO 路径计算中选择 λ 取值的算法, 以及计算组 LASSO 估计的二阶坐标下降算法; 第 3 节通过模拟数据和实际数据, 说明本文中选择 λ 值算法的可靠性, 以及相比一阶近似坐标下降算法, 二阶近似坐标下降算法的优势; 第 4 节为本文结论; 第 5 节给出文中一些关键参数的计算过程, 以及本文定理的证明.

2 广义线性模型组 LASSO 系数路径估计算法

本节将给出计算广义线性模型组 LASSO 路径 $\hat{\beta}(\lambda)$ 的算法, 有两部分内容: 首先, 推广 Park 和 Hastie^[9] 的路径算法中选择 λ 值的方法到广义线性模型组 LASSO 路径的计算中; 其次, 给出广义线性模型组 LASSO 估计的二阶近似坐标下降算法.

2.1 模型及关键变量定义

按 Fahrmeir 和 Tutz^[14] 的定义, 若 k 维变量 y 服从指数族分布, 其对数似然函数为

$$\frac{y^T \theta - b(\theta)}{\phi} + c(y, \phi),$$

其中 ϕ 是尺度参数, θ 是模型的自然参数, 与 $\mathbf{y}$ 的期望 μ 间存在函数关系 $\theta = \theta(\mu)$.

记 $\mu = (\mu_{(1)}, \dots, \mu_{(k)})^T$. 一般而言, μ 和自变量集 $\mathbf{x}$ 按如下的方式建立联系: 基于 $\mathbf{x}$ 构造 $k \times R$ 的设计矩阵 (design matrix) 变量 $\hat{Z}$; 定义线性因子 $\eta = \hat{Z}\beta$, 记 $\eta = (\eta_{(1)}, \dots, \eta_{(k)})^T$; 假定 $\mu_{(j)} = h_j(\eta_{(1)}, \dots, \eta_{(k)})$, $j = 1, \dots, k$, 简记为 $\mu = h(\eta)$. 进一步, 可以得到映射关系 $\theta = \theta(\eta)$.

记 $(\mathbf{x}, \mathbf{y})$ 的 n 组观测为 $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \dots, n\}$, 相应地, Z_i , η_i 和 μ_i 分别为 $\hat{Z}$, η 和 μ 在第 i 组观测中对应的取值. 与 Friedman 等人^[10] 的处理方式相同, 本文仅考虑对数似然函数中与模型系数相关的部分时, 广义线性模型下损失函数 $L(\beta)$ 为

$$L(\beta) = - \sum_{i=1}^n (\mathbf{y}_i^T \theta_i - b(\theta_i)), \quad (2.1)$$

其中 θ_i 是 $\theta(\eta_i)$ 的简写.

沿用 Fahrmeir 和 Tutz^[13] 的定义, 记 $Z = (Z_1^T, \dots, Z_n^T)^T$, $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T$, $\mu = (\mu_1^T, \dots, \mu_n^T)^T$,

$$D_i(\beta) = \frac{\partial h(\eta)^T}{\partial \eta} \bigg|_{\eta=\eta_i}, \quad \Sigma_i(\beta) = \frac{\partial^2 b(\theta)}{\partial \theta \partial \theta^T} \bigg|_{\eta=\eta_i}, \quad W_i(\beta) = D_i(\beta) \Sigma_i^{-1}(\beta) D_i(\beta)^T, \quad i = 1, \dots, n,$$

$$D(\beta) = \text{diag}(D_i(\beta)), \quad \Sigma(\beta) = \text{diag}(\Sigma_i(\beta)), \quad W(\beta) = \text{diag}(W_i(\beta)),$$

以及

$$\mathbf{s}(\beta) = \frac{\partial L(\beta)}{\partial \beta} = -Z^T D(\beta) \Sigma^{-1}(\beta) (\mathbf{y} - \mu), \quad F(\beta) = \frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} = Z^T W(\beta) Z.$$

不失一般性, 根据所研究问题中的组结构约束, 可将 $nk \times R$ 维的矩阵 Z 表示为 $Z = (\mathbf{z}_0, \dots, \mathbf{z}_P)$, 其中 $\mathbf{z}_p$ 为 $nk \times R_p$ 维矩阵, $\sum_{p=0}^P R_p = R$. 在 $k = 1$ 时, 例如, logistic 回归模型中, $\mathbf{z}_0 = \mathbf{1}_n$, $k > 1$ 时, 根据各类不同模型 (如离散选择模型、累积 logistic 回归模型等) 的设定确定 $\mathbf{z}_0$ (参见文献 [13]). 相应地, 模型系数可表示为 $\beta = (\beta_0^T, \dots, \beta_P^T)^T$, β_p 为 $\mathbf{z}_p$ 对应的回归系数, $p = 0, \dots, P$.

另外, 记

$$\mathbf{s}_p(\beta) = \frac{\partial L(\beta)}{\partial \beta_p} = \mathbf{z}_p^T D(\beta) \Sigma^{-1}(\beta) (\mathbf{y} - \mu), \quad F_p(\beta) = \frac{\partial^2 L(\beta)}{\partial \beta_p \partial \beta_p^T} = \mathbf{z}_p^T W \mathbf{z}_p, \quad p = 0, \dots, P.$$

一般情形下, 各类收缩方法均不惩罚其对应的系数 β_0 . Yuan 和 Lin^[2] 所定义组 LASSO 的罚函数形式为

$$T(\beta) = \sum_{p=1}^P \sqrt{\beta_p^T K_p \beta_p}, \quad (2.2)$$

其中 K_p 为 $R_p \times R_p$ 维的正定矩阵. 通常 $K_p = R_p I_{R_p}$ ($p = 1, \dots, P$), 此时, $T(\beta) = \sum_{p=1}^P \sqrt{R_p} \|\beta_p\|_2$, 不失一般性, 本文也选择这种形式. 如果 $R_p = 1$ ($p = 1, \dots, P$), 则组 LASSO 的罚函数退化为 LASSO 罚函数.

根据 (1.1)、(2.1) 和 (2.2), 定义广义线性模型组 LASSO 路径为

$$\hat{\beta}(\lambda) = \arg \min_{\beta} (\Gamma(\beta; \lambda)) = \arg \min_{\beta} - \sum_{i=1}^n (\mathbf{y}_i^T \theta_i - b(\theta_i)) + \lambda \sum_{p=1}^P \sqrt{R_p} \|\beta_p\|_2. \quad (2.3)$$

2.2 广义线性模型组 LASSO 路径估计中参数 λ 取值的选择

根据 KKT 条件 (Karush-Kuhn-Tucker conditions), 很容易得到下面的定理 (Roth 和 Fischer^[15] 的研究中也给出过类似的定理和相应的证明, 故略去证明).

定理 1 $\beta = (\beta_0, \dots, \beta_P)^T$ 为 (2.3) 极小值的充分必要条件为

(1) $s_0(\beta) = 0$;

(2) $\forall \|\beta_p\|_2 \neq 0, p = 1, \dots, P,$

$$s_p(\beta) + \lambda \frac{\sqrt{R_p} \beta_p}{\|\beta_p\|_2} = 0; \quad (2.4)$$

(3) $\forall \|\beta_p\|_2 = 0, p = 1, \dots, P, \|\beta_p(\beta)\|_2 \leq \lambda \sqrt{R_p}.$

定义如下两个映射: 活跃集 $\mathcal{A}(\lambda) = \{p : \|\hat{\beta}_p(\lambda)\|_2 \neq 0, p = 0, \dots, P\}$, 非活跃集 $\mathcal{A}^c(\lambda) = \{p : \|\hat{\beta}_p(\lambda)\|_2 = 0, p = 0, \dots, P\}$. 定理 1 表明, 存在 $\lambda_{\max}$, 当 $\lambda \geq \lambda_{\max}$ 时, $\mathcal{A}(\lambda) = \{0\}$; 排除极端的情形, 当 $\lambda = 0$ 时, $\mathcal{A}(\lambda) = \{0, \dots, P\}$; 当 λ 由 $\lambda_{\max}$ 向 0 变动的过程中, 各变量组会以不同的顺序进入或退出模型, 即活跃集 $\mathcal{A}(\lambda)$ 发生改变. 基于此, 类似文献 [9] 中 λ 选择的思路, 我们始于 $\lambda_{\max}$, 结束于 0, 寻找使得 $\mathcal{A}(\lambda)$ (或者 $\mathcal{A}^c(\lambda)$) 恰好改变的值. 在实际迭代算法中, 关键步骤为, 给定 $\lambda^{(m)} \in (0, \lambda_{\max})$, 找到 $h^{(m)}$, 使得

$$\lambda \in (\lambda^{(m)} - h^{(m)}, \lambda^{(m)}), \quad \mathcal{A}(\lambda) = \mathcal{A}(\lambda^{(m)}), \quad \mathcal{A}(\lambda^{(m)}) \neq \mathcal{A}(\lambda^{(m)} - h^{(m)}).$$

换言之, $\lambda = \lambda^{(m)} - h^{(m)}$ 时, 恰好有变量组退出或者进入模型. 在排除极端的情形下, 对应两种情形:

情形 1 恰好有变量组进入模型, 根据定理 1(2), 对于 $p \in \mathcal{A}^c$,

$$\|s_p(\hat{\beta}(\lambda^{(m)} - h_p^{(m)}))\|_2 = (\lambda^{(m)} - h_p^{(m)})\sqrt{R_p}; \quad (2.5)$$

情形 2 恰好有变量组退出模型, 即对于 $p \in \mathcal{A}$,

$$\|\hat{\beta}_p(\lambda^{(m)} - h_p^{(m)})\|_2 = 0. \quad (2.6)$$

类似 Park 和 Hastie^[9] 的做法, 可以按如下方法计算 $h^{(m)}$ 的近似解:

情形 1 对于 $p \in \mathcal{A}^c$, 根据 Taylor 近似,

$$s_p(\hat{\beta}(\lambda^{(m)} - h_p^{(m)})) \approx s_p(\hat{\beta}(\lambda^{(m)})) - h_p a_p(\lambda^{(m)}), \quad (2.7)$$

其中 $a_p(\lambda^{(m)}) = \partial s_p(\hat{\beta}(\lambda))/\partial \lambda|_{\lambda=\lambda^{(m)}}$. 根据 (2.5) 和 (2.7), 可以得到 h_p 的一个正数解 (具体表达式及推导过程见第 5.1 小节).

情形 2 对于 $p \in \mathcal{A}$, 根据 Taylor 近似, 可得

$$\hat{\beta}_p(\lambda^{(m)} - h_p^{(m)}) \approx \hat{\beta}_p(\lambda^{(m)}) - h_p^{(m)} \frac{\partial \hat{\beta}_p(\lambda)}{\partial \lambda} \Big|_{\lambda=\lambda^{(m)}}. \quad (2.8)$$

根据 (2.6) 和 (2.8), 可以得到 h_p 的一个正数解 (具体表达式及推导过程见第 5.1 小节).

记 $h^{(m)} = \min\{h_1^{(m)}, \dots, h_P^{(m)}\}$, $\lambda^{(m)} - h^{(m)}$ 就是当 λ 自 $\lambda^{(m)}$ 开始变小时, 活跃集恰好改变的最大值. 由此, 可以给出计算广义线性模型组 LASSO 路径中选择 λ 值的算法, 具体过程见算法 1.

算法 1 广义线性模型组 LASSO 路径中选择 λ 值的算法.

第 1 步 初始化 $m = 1$, 计算 $\lambda^{(1)} = \lambda_{\max}$ (具体计算见第 5.2 小节);

第 2 步 计算 $\hat{\beta}(\lambda^{(m)})$;

第 3 步 计算 λ 小于 $\lambda^{(m)}$ 时, 使得当前活跃集改变的最大取值 $\lambda^{(m+1)} = \lambda^{(m)} - h^{(m)}$;

第 4 步 更新迭代次数 $m = m + 1$, 重复第 2 和第 3 步, 直至所有变量组进入模型, 停止算法.

由于在 $h^{(m)}$ 的计算过程中使用了一系列的近似, 可能会出现步长 h 过大的情形, 使得不能准确地找到对应活跃集改变的关键 λ 值. 为克服这个问题, 考虑在算法 1 的第 3 步之后追加一个调整步骤: 当 p 进入或者退出活跃集时, 若 $\|\beta_p\|_2 \geq \delta$ (δ 为任意小的一个正数), 将 $h^{(m)}$ 适当调小直至新进入组变量系数模小于给定范围为止 ($\|\beta_p\|_2 < \delta$), 在本文的计算程序中选取了比较保守的做法, 从 $\{0.9^l h, l = 1, 2, \dots\}$ 中选择满足条件的最大元素. 关于 δ , 本文使用 $R_p \times 10^{-3}$, 其中 R_p 为对应组变量的自由度.

2.3 估计 $\hat{\beta}(\lambda^{(m)})$ 的二阶近似坐标下降算法

给定 $\lambda = \lambda^{(m)}$ 情形下, 为了方便表述, 在不至混淆的情形下, 简记组 LASSO 的解 $\hat{\beta}(\lambda^{(m)})$ 为 $\hat{\beta}$. 在坐标下降算法计算 $\hat{\beta}$ 的过程中, 核心步骤是构造下降方向.

文献 [3, 12] 利用 $L(\beta)$ 的一阶导 $s(\beta)$, 构造的下降方向为

$$d_H(\beta; p) = \arg \min_d \left\{ (s(\beta))^T d + \frac{1}{2} d^T H d + \lambda T(\beta + d) \mid d_j = 0, \forall j \neq p \right\}, \quad p = 1, \dots, P, \quad (2.9)$$

其中 $d = (d_0^T, \dots, d_P^T)^T$, d_p 为 R_p 维向量, 是 β_p 对应的方向, H 是 $R \times R$ 维的分块对角阵, 其第 j 个对角子矩阵为

$$H_j = \begin{cases} k_j I_{R_j}, & j = p, \\ \mathbf{0}_{R_j}, & j \neq p, \end{cases}$$

这里 $k_j = \min\{\max\{\text{diag}(F_j(\beta)), 10^{-2}\}, 10^9\}$.

类似地, 利用 $L(\beta)$ 的一阶导 $s(\beta)$ 和二阶导 $F(\beta)$ 构造下降方向

$$d_F(\beta; p) = \arg \min_d \left\{ (s(\beta))^T d + \frac{1}{2} d^T F(\beta) d + \lambda T(\beta + d) \mid d_j = 0, \forall j \neq p \right\}, \quad p = 1, \dots, P, \quad (2.10)$$

其中 $d = (d_0^T, \dots, d_P^T)^T$, 可以得到以下定理.

定理 2 记 $d_H = d_H(\beta; p)$, $d_F = d_F(\beta; p)$, $s = s(\beta)$, $F = F(\beta)$, 对于任意 β ,

$$\Gamma(\beta + \alpha d_H) \leq \Gamma(\beta) + \alpha \Delta_H + o(\alpha), \quad (2.11)$$

$$\Gamma(\beta + \alpha d_F) \leq \Gamma(\beta) + \alpha \Delta_F + o(\alpha^2), \quad (2.12)$$

其中

$$\begin{aligned} \Delta_H &= s^T d_H + \lambda(T(\beta + d_H) - T(\beta)) \leq -\frac{1}{2}(1 - \gamma) d_H^T H d_H, \\ \Delta_F &= s^T d_F + \frac{1}{2} \alpha d_F^T F d_F + \lambda(T(\beta + d_F) - T(\beta)) \leq -\frac{1}{2}(1 - \gamma) d_F^T F d_F, \end{aligned}$$

这里 $0 \leq \gamma < 1$.

(2.11) 的证明可以根据文献 [11, 引理 2.1] 得到, (2.12) 的证明见第 5.3 小节. 定理 2 中, (2.11) 中的 $o(\alpha)$ 是步长 α 的高阶无穷小, 而 (2.12) 中的 $o(\alpha^2)$ 是步长 α^2 的高阶无穷小, 使用 Armijo 规则, 二

阶近似坐标下降算法对应的步长很可能比一阶近似坐标下降算法对应的步长大; 这意味着二阶近似的坐标下降算法中, 搜索步长的时间可能更少, 目标函数的下降可能更大.

尽管二阶近似坐标下降算法中, 下降方向 $\mathbf{d}_F(\boldsymbol{\beta}; p)$ 没有显式解, 但很容易使用数值算法求解. 记 $\mathbf{d}_F(\boldsymbol{\beta}; p) = ((\mathbf{d}_{F,0}(\boldsymbol{\beta}; p))^T, \dots, (\mathbf{d}_{F,P}(\boldsymbol{\beta}; p))^T)^T$, 其中 $\mathbf{d}_{F,j}(\boldsymbol{\beta}; p)$ 为 R_j 维向量. 为便于表达, 记 $\mathbf{s}_p = \mathbf{s}_p(\boldsymbol{\beta})$, $F_p = F_p(\boldsymbol{\beta})$, $\mathbf{d}_{F,p} = \mathbf{d}_{F,p}(\boldsymbol{\beta}; p)$, 定义 $\mathbf{b}_p = \sqrt{R_p}(\boldsymbol{\beta}_p + \mathbf{d}_p)$, $B_p = F_p/R_p$, $\mathbf{g}_p = (\mathbf{s}_p - F_p^T \boldsymbol{\beta}_p)/\sqrt{R_p}$, (2.12) 等价于

$$\mathbf{b}_{F,p} = \min_{\mathbf{b}_p} \mathbf{g}_p^T \mathbf{b}_p + \frac{1}{2} \mathbf{b}_p^T B_p \mathbf{b}_p + \lambda \|\mathbf{b}_p\|_2. \quad (2.13)$$

由于 $B_p \succeq 0$, 根据 KKT 条件, 可以得到如下定理.

定理 3 $\mathbf{b}_p$ 为 (2.13) 解的充分必要条件为

$$\mathbf{b}_p = \begin{cases} \mathbf{0}_{R_p \times 1}, & \text{如果 } \|\mathbf{g}_p\|_2 \leq \lambda, \\ -\left(B_p + \frac{\lambda}{t_p} I_{R_p}\right)^{-1} \mathbf{g}_p, & \text{其中 } \|\mathbf{b}_p\|_2 = t_p, \text{ 如果 } \|\mathbf{g}_p\|_2 > \lambda. \end{cases}$$

Qin 等人^[16] 在阐述 BCD-GL 算法 (Block coordinate descent (BCD) algorithms for group LASSO) 时以信任域 (trust region) 理论为工具, 给出了类似的定理及相应的证明. 根据定理 3, 可以有

$$\|(t_p B_p + \lambda I_{R_p})^{-1} \mathbf{g}_p\|_2^2 = 1. \quad (2.14)$$

基于 (2.14), 可以使用 Newton 迭代法计算 t_p (t_p 的计算细节见第 5.4 小节). 一般情形下, R_p 的值都较小, 相应的 Newton 迭代法可以很快收敛.

基于定理 2 和 3, 可以给出二阶近似坐标下降算法如下.

算法 2 二阶近似坐标下降算法.

第 1 步 设定 $\boldsymbol{\beta}$ 的初始值 $\boldsymbol{\beta}_F^{(0)} = \mathbf{0}$, 迭代步数 $m = 0$;

第 2 步 根据 Tseng 和 Yun^[11] 的定义, 按照 Gauss-Seidel 规则设定 $p^{(m)}$, 计算 $\mathbf{d}_F(\boldsymbol{\beta}^{(m)}; p^{(m)})$;

第 3 步 根据 Armijo 规则计算步长 $\alpha^{(m)}$;

第 4 步 更新 $\boldsymbol{\beta}_F^{(m+1)} = \boldsymbol{\beta}_F^{(m)} + \alpha^{(m)} \mathbf{d}_F(\boldsymbol{\beta}_F^{(m)}; p^{(m)})$;

第 5 步 更新迭代次数 $m = m + 1$, 重复第 1 至第 4 步, 当 m 足够大或 $\sum_{m=m^*}^{m^*+P} \|\mathbf{d}_F(\boldsymbol{\beta}^{(m)}; p^{(m)})\| < \epsilon$ ($m > m^*$, ϵ 为任意小常数) 条件满足时, 停止算法.

与文献 [11] 的一阶近似坐标下降算法类似, 基于二阶近似的坐标下降算法满足收敛性, 有如下定理成立.

定理 4 若存在常数 $\bar{\delta}$ 和 $\underline{\delta}$ 使得 $\bar{\delta} I \succeq F^{(m)} \succeq \underline{\delta} I$, 其中 $\bar{\delta} \geq \underline{\delta} \geq 0$, 则 $\mathbf{d}_F(\boldsymbol{\beta}^{(m)}; p^{(m)})$ 和 $\boldsymbol{\beta}_F^{(m)}$ 具有如下极限性质:

$$\lim_{m \rightarrow +\infty} \mathbf{d}_F(\boldsymbol{\beta}^{(m)}; p^{(m)}) = \mathbf{0}_P, \quad (2.15)$$

$$\lim_{m \rightarrow +\infty} \boldsymbol{\beta}_F^{(m)} = \hat{\boldsymbol{\beta}}. \quad (2.16)$$

也就是说, 算法 2 中, 只要 m 足够大, 第 5 步的停止条件总能满足. 定理 4 的证明见第 5.5 小节.

3 模拟和实际数据分析

本节通过模拟数据和实际数据, 验证第 2.2 小节中算法 1 在选择 λ 值的可靠性, 分析第 2.3 小节中的二阶近似坐标下降算法较一阶近似坐标下降算法在速度等方面的优势.

3.1 模拟数据分析: 路径计算中选择 λ 值算法的可靠性

数据模拟方式与文献 [2] 类似, 具体过程如下.

数据 1 记 5 维正态分布 $(U_1, \dots, U_5) \sim N_5(\mathbf{0}, \Sigma)$, 其中 $\Sigma_{m,n} = 0.5^{|m-n|}$, $m, n \in \{1, 2, \dots, 5\}$, 根据 U_j 定义分类变量 X_j , 其第 i 个观测依次记为 U_{ji} 和 X_{ji} , 令

$$X_{ji} = \begin{cases} 1, & U_{ji} \leq \Phi^{-1}(0.2), \\ 2, & \Phi^{-1}(0.2) < U_{ji} \leq \Phi^{-1}(0.5), \\ 3, & \Phi^{-1}(0.5) < U_{ji} \leq \Phi^{-1}(0.9), \\ 4, & U_{ji} > \Phi^{-1}(0.9), \end{cases}$$

其中 $i = 1, \dots, 100$, $j = 1, \dots, 5$.

根据 X_{ji} , 可定义 1×3 维的设计向量 $\mathbf{z}_{ji}$. 记

$$\eta_i = \beta_0 + \mathbf{z}_{1i}\beta_1 + \mathbf{z}_{3i}\beta_3 + \mathbf{z}_{4i}\beta_4,$$

其中 $i = 1, \dots, 100$, $\beta_0, \beta_1, \beta_3, \beta_4$ 中的元素为服从 $U(-1, 1)$ 分布的随机数.

使用 logistic 函数作为连接函数, 即 $\mu_i = \exp(\eta_i)/(1 + \exp(\eta_i))$. 记因变量 y 的第 i 个模拟值为 y_i , $y_i \sim \text{Bernoulli}(1, \mu_i)$.

算法 1 经过 21 次迭代之后, 5 组变量全部进入模型. 为了考察其可靠性, 从区间 $(0, \lambda^{(1)}]$ 内等间隔取 900 个点, 在每一个 λ 上也使用二阶近似算法估计了自变量系数. 由于格点法中, λ 取值极为密集, 可将其看作真实系数路径. 表 1 给出了算法 1 和格点法分别选出的恰好对应活跃集改变的 λ 值的对比. 二者找到的 λ 值非常接近, 差的绝对值在 0.005 以内. 两个算法所使用的时间基本与迭代次数是成正比的, 格点法为 2915 秒, 算法 1 为 85 秒. 与格点法相比, 算法 1 使用较少的计算时间达到了选择使得变量恰好进入模型的关键 λ 值.

3.2 模拟数据分析: 二阶近似算法与一阶近似算法的对比

本节将考虑以下 4 组模拟数据比较本文所提出的二阶近似算法和 Meier 等人^[3]所使用的一阶近似算法在给定参数 λ 情形下求解组 LASSO 估计时的表现.

数据 2 首先模拟一个 9 维正态分布 $(U_1, \dots, U_9) \sim N_9(\mathbf{0}, \Sigma)$, 其中 $\Sigma_{i,j} = 0.5^{|i-j|}$, $i, j \in \{1, \dots, 9\}$. 根据 U_j 定义分类变量 X_j , 其第 i 个观测依次记为 U_{ji} 和 X_{ji} , 令

$$X_{ji} = \begin{cases} 1, & U_{ji} \leq \Phi^{-1}(0.2), \\ 2, & \Phi^{-1}(0.2) < U_{ji} \leq \Phi^{-1}(0.5), \\ 3, & \Phi^{-1}(0.5) < U_{ji} \leq \Phi^{-1}(0.9), \\ 4, & U_{ji} > \Phi^{-1}(0.9), \end{cases}$$

其中 $i = 1, \dots, 500$, $j = 1, \dots, 9$.

根据 X_{ji} , 定义 1×3 维的设计向量 $\mathbf{z}_{ji}$ ($j = 1, \dots, 9$), 记

$$\eta_i = \beta_0 + \sum_{j=1}^9 \mathbf{z}_{ji}\beta_j,$$

表 1 算法 1 与格点法的迭代次数以及所寻找到的活跃集改变对应 λ 值对比

	迭代次数	λ_1	λ_2	λ_3	λ_4	λ_5
算法 1	21	5.377046	2.989572	2.866610	2.555066	1.219578
格点法	900	5.377046	2.987479	2.862024	2.551373	1.219159

其中 $i = 1, \dots, 500$, β_0 以及 $\beta_1, \dots, \beta_6$ 中的元素为服从 $U(-1, 1)$ 分布的随机数.

与数据 1 类似, 使用 logistic 函数作为连接函数并得到 μ_i , 因变量 y 的第 i 个观测 y_i 服从参数为 μ_i 的两点分布.

数据 3 在数据 2 基础上, 数据 3 将自变量集扩大到 29 组 3 水平的变量和 1 个截距项, 其中只有前 20 组 5 水平变量和截距项对应的模型系数不为 0.

数据 4 记 9 维正态分布 $(U_1, \dots, U_9) \sim N_9(\mathbf{0}, \Sigma)$, 其中 $\Sigma_{i,j} = 0.5^{|i-j|}$, $i, j \in \{1, \dots, 9\}$.

根据 U_j 定义分类变量 X_j , 其第 i 个观测依次记为 U_{ji} 和 X_{ji} , 令

$$X_{ji} = \begin{cases} 1, & U_{ji} \leq \Phi^{-1}(0.15), \\ 2, & \Phi^{-1}(0.15) < U_{ji} \leq \Phi^{-1}(0.5), \\ 3, & \Phi^{-1}(0.5) < U_{ji} \leq \Phi^{-1}(0.7), \\ 4, & \Phi^{-1}(0.7) < U_{ji} \leq \Phi^{-1}(0.85), \\ 5, & \Phi^{-1}(0.85) < U_{ji} \leq \Phi^{-1}(0.9), \\ 6, & U_{ji} > \Phi^{-1}(0.9), \end{cases}$$

其中 $i = 1, \dots, 500$, $j = 1, \dots, 9$. 在模型中描述 X_j ($j = 1, \dots, 9$) 时需要使用 5 个虚拟变量, 因此, 该数据的自变量集共包括 9 组水平为 5 ($df = 5$) 的变量和 1 个截距项.

根据 X_{ji} , 可定义 1×5 维的设计向量 $\mathbf{z}_{ji}$ ($j = 1, \dots, 9$), 记

$$\eta_i = \beta_0 + \sum_{j=1}^6 \mathbf{z}_{ji} \beta_j$$

的元素为服从 $U(-1, 1)$ 分布的随机数.

同样使用 logistic 函数作为连接函数构造 μ_i , 因变量 y 的第 i 个观测 y_i 服从参数为 μ_i 的两点分布.

数据 5 在数据 4 基础上, 数据 5 将自变量集扩大到 29 组 5 水平变量和 1 个截距项, 其中只有前 20 组 5 水平变量和截距项对应的模型系数不为 0.

本文在实际计算过程中设定循环次数固定为 200 次, 若算法在这 200 次迭代中还未达到收敛状态, 也将结束计算. 表 2 给出了 $\lambda \in \{1, 2, \dots, 15\}$ 时, 二阶近似坐标下降算法和一阶近似坐标下降算法分别完成 200 次循环的平均运行时间 (本文的运行时间为, 在 CPU 为 Core2 双核 T9400, 内存为 2GB DDR 1066MHZ 的笔记本电脑上, Win7 32 位系统环境下, 由 Inter(R) Visual Fortran 11.1.065[IA-32] 编译器编译 Fortran95 语言程序得到的可执行程序的计算时间). 表 2 中的结果表明, 二阶近似算法完成限定次数循环所使用的时间整体少于一阶近似算法, 说明前者的平均每次迭代速度快于后者, 而且, 自变量集维数越高时, 两者在计算时间上的差距越大.

表 3 给出了 4 组模拟数据在 $\lambda \in \{2k; k = 1, 2, \dots, 7\}$ 时, 给定循环 200 次的情形下, 所得目标函

表 2 限定循环次数时二阶近似坐标下降算法与一阶近似坐标下降算法的平均运行时间对比 (单位: 秒)

	数据 2	数据 3	数据 4	数据 5
二阶近似	8.9367	123.6937	16.4568	217.8555
一阶近似	14.0891	165.7242	21.7776	303.9755

表 3 限定循环次数时二阶坐标下降近似算法与一阶近似坐标下降算法所达到的目标函数值对比

λ	数据 2		数据 3		数据 4		数据 5	
	二阶近似	一阶近似	二阶近似	一阶近似	二阶近似	一阶近似	二阶近似	一阶近似
2	318.2275	318.2279	211.2704	241.3949	292.9350	292.9351	257.2557	257.2559
4	324.1365	328.8848	263.0521	279.8235	312.2697	312.2697	295.2043	295.2045
6	332.7007	335.7982	293.0901	303.7727	324.7971	324.7970	310.9431	310.9432
8	338.3098	340.3387	312.3860	319.3047	332.9522	332.9521	318.3036	318.3035
10	342.0081	343.3482	324.7607	329.0620	337.7260	337.7259	322.6677	322.6677
12	344.3689	345.0773	332.3542	334.7672	339.6411	339.6410	324.9290	324.9290
14	345.5243	345.7579	336.4609	337.6225	339.8192	339.8192	325.5364	325.5362

数值的详细对比. 4 组数据中, 二阶近似坐标下降算法较之一阶近似坐标下降算法均能得到相近或更低的目标函数值, 且当数据维数较高时, 二阶近似坐标下降算法的优势更显著.

3.3 实际数据分析

本文所使用的实际数据来自国家统计局的“中国城镇居民经济状况与心态调查”项目 (该项目旨在调查及研究中国的自有住房状况与国民幸福感之间的关系, 详见文献 [17]), 与本文分析有关的变量描述见表 4, 关于该数据集中各组变量的详细经济含义等其他信息请参见文献 [17].

因变量为“是否幸福”, 自变量集为表 4 中其余 7 组哑变量. 算法 1 经过 55 次迭代将全部 7 组变量引入模型, 与模拟研究中使用的方法类似, 从区间 $(0, \lambda^{(1)}]$ 内等间隔抽取 300 个 λ 值, 并计算了相应的组 LASSO 估计. 算法 1 和格点法得到的使 7 组变量恰好进入模型的 λ 值见表 5, 二者差的绝对值在 0.006 以内.

当固定 $\lambda \in \{3, 2.5, 2, 1.5, 1, 0.5\}$ 时, 本文使用该数据也对比了一阶近似算法和二阶近似算法在固定 200 次循环的情形下所最终达到的目标函数值, 其结果如表 6 所示. 平均而言, 二阶近似算法的运行时间为 4.9831 秒, 而一阶算法的运行时间为 7.6957 秒. 在最终达到的目标函数值接近的前提下, 二阶近似算法的速度优势比较明显.

表 4 幸福感数据变量说明

变量名	组内变量数	变量名	组内变量数
是否幸福	1	家庭人均年收入	3
健康状况	2	非住房家庭资产	3
是否有大产权住房	1	是否有小产权住房	1
是否有非住房类负债	1	性别	1

表 5 幸福感数据: 算法 1 与格点法的迭代次数以及所寻找到的活跃集改变对应 λ 值对比

	迭代次数	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
算法 1	55	2.9538	2.4123	1.7624	1.1420	1.1321	0.4724	0.0589
格点法	300	2.9538	2.4160	1.7622	1.1477	1.1324	0.4766	0.0607

表 6 幸福感数据: 限定循环次数时二阶坐标下降近似算法与一阶近似坐标下降算法所达到的目标函数值对比

λ	3.0	2.5	2.0	1.5	1.0	0.5
二阶近似	41.9859	41.9257	41.3173	39.9096	37.6460	34.2620
一阶近似	41.9859	41.9257	41.3120	39.9136	37.6495	34.2655

综合来看, 对于给点 λ 的组 LASSO 估计, 相比一阶近似算法, 二阶近似算法的收敛速度更快; 而算法 1 可以用 55 次迭代, 刻画自变量集进入模型的顺序: “是否用拥有大产权住房”、“家庭人均收入”、“健康状况”、“是否有非住房类负债”、“非住房家庭资产”、“是否拥有小产权住房”、“性别”, 相比格点法, 计算量更小. 在此基础上, 只需要 7 次交叉验证或者 AIC (Akaike information criterion) 等模型选择指标, 就可以选择影响因变量的因素.

4 结论

本文在文献 [9] 的基础上, 给出了广义线性模型组 LASSO 路径中 λ 值的选择算法, 并利用文献 [12] 中坐标下降算法的思想, 提出了基于对广义线性模型损失函数进行二阶 Taylor 近似的坐标下降算法, 可以更高效地求解固定调节参数 λ 情形下的组 LASSO 估计问题. 模拟数据和实际数据的分析结果表明, 本文所提出的路径参数值选择算法使用较少的 λ 值, 可以相对准确地发现路径中的变点 (活跃集改变的点). 另外, 本文所使用的二阶近似坐标下降算法与文献 [3, 12] 所使用的一阶近似坐标下降算法相比, 具有收敛速度快、效率高等优良性质.

5 关键变量的计算细节及定理的证明

5.1 $h_p^{(m)}$ 的计算细节

不失一般性, 假定 $\mathcal{A}(\lambda) = \{p_1, \dots, p_Q\}$, 记 $\mathbf{z}_{\mathcal{A}} = (\mathbf{z}_{p_1}^T, \dots, \mathbf{z}_{p_Q}^T)^T$, $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda) = (\boldsymbol{\beta}_{p_1}^T, \dots, \boldsymbol{\beta}_{p_Q}^T)^T$,

$$\mathbf{c}_{\mathcal{A}}(\boldsymbol{\beta}, \lambda) = \frac{\partial \Gamma(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}_{\mathcal{A}}} = \mathbf{z}_{\mathcal{A}}^T D(\boldsymbol{\beta}) \Sigma^{-1}(\boldsymbol{\beta}) (\mathbf{y} - \boldsymbol{\mu}) + \lambda \mathbf{u}_{\mathcal{A}}(\boldsymbol{\beta}),$$

其中

$$\mathbf{u}_{\mathcal{A}}(\boldsymbol{\beta}) = \left(\left(\sqrt{R_{p_1}} \frac{\boldsymbol{\beta}_{p_1}}{\|\boldsymbol{\beta}_{p_1}\|_2} \right)^T, \dots, \left(\sqrt{R_{p_Q}} \frac{\boldsymbol{\beta}_{p_Q}}{\|\boldsymbol{\beta}_{p_Q}\|_2} \right)^T \right)^T,$$

$\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda)$ 实质是上述方程决定的隐函数.

另外, 记

$$\mathbf{V}_{\mathcal{A}}(\boldsymbol{\beta}; \lambda) = \frac{\partial^2 \Gamma(\boldsymbol{\beta}; \lambda)}{\partial \boldsymbol{\beta}_{\mathcal{A}} \partial \boldsymbol{\beta}_{\mathcal{A}}^T} = \mathbf{Z}_{\mathcal{A}}^T \mathbf{W}(\boldsymbol{\beta}) \mathbf{Z}_{\mathcal{A}} + \lambda \mathbf{v}_{\mathcal{A}}(\boldsymbol{\beta}),$$

其中

$$v_{\mathcal{A}}(\beta) = \begin{pmatrix} \sqrt{R_{p_1}} \left(\frac{I_{R_{p_1}}}{\|\beta_{p_1}\|_2} - \frac{\beta_{p_1} \beta_{p_1}^T}{\|\beta_{p_1}\|_2^3} \right) & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \sqrt{R_{p_Q}} \left(\frac{I_{R_{p_Q}}}{\|\beta_{p_Q}\|_2} - \frac{\beta_{p_Q} \beta_{p_Q}^T}{\|\beta_{p_Q}\|_2^3} \right) \end{pmatrix}.$$

记

$$\mathbf{a}_p(\lambda) = \frac{\partial \mathbf{s}_p(\beta)}{\partial \lambda} = \frac{\partial \mathbf{s}_p(\beta)}{\hat{\beta}_{\mathcal{A}}(\lambda)} \frac{\hat{\beta}_{\mathcal{A}}(\lambda)}{\partial \lambda},$$

其中

$$\frac{\partial \mathbf{s}_p(\beta)}{\hat{\beta}_{\mathcal{A}}(\lambda)} = -\mathbf{z}_p^T W \mathbf{z}_{\mathcal{A}}.$$

根据隐函数定理,

$$\begin{aligned} \frac{\partial \mathbf{c}_{\mathcal{A}}(\beta, \lambda)}{\partial \lambda} &= \frac{\partial \mathbf{c}_{\mathcal{A}}(\beta, \lambda)}{\partial \lambda} + \frac{\partial \mathbf{c}_{\mathcal{A}}(\beta, \lambda)}{\partial \beta_{\mathcal{A}}} \frac{\partial \beta_{\mathcal{A}}}{\partial \lambda} = 0, \\ \frac{\partial \hat{\beta}_{\mathcal{A}}(\lambda)}{\partial \lambda} &= - \left(\frac{\partial \mathbf{c}_{\mathcal{A}}(\beta, \lambda)}{\partial \beta_{\mathcal{A}}} \right)^{-1} \frac{\partial \mathbf{c}_{\mathcal{A}}(\beta, \lambda)}{\partial \lambda} = -(V_{\mathcal{A}}(\beta; \lambda))^{-1} \mathbf{u}_{\mathcal{A}}, \end{aligned}$$

因而,

$$\mathbf{a}_p(\lambda) = -(\mathbf{z}_p^T W(\beta) \mathbf{z}_{\mathcal{A}}) (V_{\mathcal{A}}(\beta; \lambda))^{-1} \mathbf{u}_{\mathcal{A}}. \quad (5.1)$$

为便于表述, 记 $\mathbf{a}_p = \mathbf{a}_p(\lambda)$, $\mathbf{s}_p = \mathbf{s}_p(\hat{\beta}(\lambda^{(m)}) - h_p^{(m)})$. 根据定理 1, 可得

$$\|\mathbf{s}_p - h_p^{(m)} \mathbf{a}_p\|_2^2 = (\lambda^{(m)} - h_p^{(m)})^2 R_p. \quad (5.2)$$

方程 (5.2) 是一个关于 $h_p^{(m)}$ 二元一次方程, 可设函数

$$f(h_p^{(m)}) = \|\mathbf{s}_p - h_p^{(m)} \mathbf{a}_p\|_2^2 - (\lambda^{(m)} - h_p^{(m)})^2 R_p.$$

当 $h_p^{(m)} = 0$ 时, $f(h_p^{(m)}) < 0$; 当 $h_p^{(m)} = \lambda^{(m)}$ 时, $f(h_p^{(m)}) \geq 0$. 因此, 方程 (4.4) 必至少有一个根落在 $(0, \lambda^{(m)})$ 区间内, 令

$$h_p^{(m)} = \min^+ \left(\frac{(\mathbf{s}_p^T \mathbf{a}_p - \lambda^{(m)} R_p) \pm \sqrt{(\mathbf{s}_p^T \mathbf{a}_p - \lambda^{(m)} R_p)^2 + (\mathbf{a}_p^T \mathbf{a}_p - R_p)(\lambda^{(m)2} R_p - \mathbf{s}_p^T \mathbf{s}_p)}}{(\mathbf{a}_p^T \mathbf{a}_p - R_p)} \right),$$

其中 $\min^+(a, b)$ 表示取 a 和 b 中为正的数中较小的一个.

对于 $p \in \mathcal{A}$, 根据 (2.8),

$$h_p^{(m)} = \frac{\|\hat{\beta}_p(\lambda^{(m)})\|_2}{\|(V_{\mathcal{A}}(\hat{\beta}(\lambda^{(m)})); \lambda^{(m)})^{-1} \mathbf{u}_{\mathcal{A}}(\hat{\beta}(\lambda^{(m)}))\|_2}.$$

5.2 λ 初值的选取方法

首先将 $\beta_0^{(0)}$ 的初值选为 $\mathbf{g}(\bar{\mathbf{y}})$, $\beta_p^{(0)} = \mathbf{0}$ ($p \in 1, \dots, P$), 令

$$q = \arg \max_{p=1, \dots, P} \|Z_p^T(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1})\|,$$

此时, 使得 $\mathcal{A} = \{0, q\}$, 使用第 2.3 小节中的二阶近似算法求解组 LASSO, 其所得解记为 $\beta^{(1)}$, 令

$$\lambda^{(1)} = \frac{\|Z_q^T W_{\mathcal{A}}(\mathbf{y} - \boldsymbol{\mu}) \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}\|_2}{\sqrt{R_q}}.$$

5.3 定理 2 的证明

证明 $\forall \alpha \in (0, 1]$, 根据 $\Gamma(\cdot)$ 的定义可知,

$$\Gamma(\beta + \alpha \mathbf{d}_F) - \Gamma(\beta) = L(\beta + \alpha \mathbf{d}_F) - L(\beta) + \lambda T(\beta + \alpha \mathbf{d}_F) - \lambda T(\beta).$$

由于 $\mathbf{s}(\beta) = \nabla L(\beta)$, $\mathbf{F}(\beta) = \nabla^2 L(\beta)$, 将 $L(\beta + \alpha \mathbf{d}_F)$ 和 $L(\beta)$ 二阶 Taylor 展开,

$$\begin{aligned} \Gamma(\beta + \alpha \mathbf{d}_F) - \Gamma(\beta) &\leq \alpha \mathbf{s}^T \mathbf{d}_F + \frac{\alpha^2}{2} \mathbf{d}_F^T \mathbf{F} \mathbf{d}_F + o(\alpha^2) + \lambda T[\alpha(\beta + \mathbf{d}_F) + (1 - \alpha)\beta] - \lambda T(\beta) \\ &= \alpha \Delta_F + o(\alpha^2), \end{aligned}$$

其中

$$\Delta_F = \mathbf{s}^T \mathbf{d}_F + \frac{1}{2} \alpha \mathbf{d}_F^T \mathbf{F} \mathbf{d}_F + \lambda(T(\beta + \mathbf{d}_F) - T(\beta)),$$

即

$$\Gamma(\beta + \alpha \mathbf{d}_F) \leq \Gamma(\beta) + \alpha \Delta_F + o(\alpha^2).$$

证毕. □

5.4 t_p 的计算细节

定理 3 给出了基于二阶近似的分块坐标下降方向 $\mathbf{d}_F(\beta; p)$ 的原理, t_p 的计算细节如下.

记 $\Upsilon(t_p) = (t_p B_p + \lambda I_{R_p})^{-1} \mathbf{g}_p$, 则 t_p 是 $\|\Upsilon(t_p)\|_2^2 = 1$ 的根. 记

$$\phi(t_p) = \|\Upsilon(t_p)\|_2^2 - 1.$$

利用 Newton 迭代法求解该方程. 记 $t_p^{(l)}$ 第 l 次迭代中的 t_p 值,

$$t_p^{(l+1)} = t_p^{(l)} - \frac{\phi(t_p^{(l)})}{\phi'(t_p^{(l)})}.$$

因此, 首先要求得 $\phi'(t_p)$. 另外, 为保证 Newton 迭代法的收敛性, 需观察 $\phi''(t_p)$.

由于 B_p 为非负定矩阵, 所以, B_p 存在的谱分解使得 $B_p = Q \Lambda Q^T$, 其中 Q 是矩阵 B_p 的 R_p 个彼此正交的单位特征向量按列组成的矩阵, $\Lambda = \text{diag}(\gamma_1, \dots, \gamma_{R_p})$, γ_i 为 B_p 的 R_p 个特征根 ($\gamma_1 \geq \dots \geq \gamma_{R_p} \geq 0$). 代入 $\Upsilon(t_p)$ 可得

$$\Upsilon(t_p) = -Q(t_p \Lambda + \lambda I_{R_p})^{-1} Q^T \mathbf{g}_p = -\sum_i^{R_p} \frac{\mathbf{q}_i^T \mathbf{g}_p}{\gamma_i t_p + \lambda} \mathbf{q}_i, \quad \phi(t_p) = \sum_i^{R_p} \frac{(\mathbf{q}_i^T \mathbf{g}_p)^2}{(\gamma_i t_p + \lambda)^2} - 1.$$

所以,

$$\phi'(t_p) = -2 \sum_i^{R_p} \frac{(\mathbf{q}_i^T \mathbf{g}_p)^2 \gamma_i}{(\gamma_i t_p + \lambda)^3} < 0, \quad \phi''(t_p) = 6 \sum_i^{R_p} \frac{(\mathbf{q}_i^T \mathbf{g}_p)^2 \gamma_i^2}{(\gamma_i t_p + \lambda)^4} > 0.$$

$\phi(t_p)$ 的一阶导和二阶导均有保号性, 因此使用 Newton 迭代法所得序列 $t_p^{(l)}$ ($l = 1, 2, \dots$) 将单调收敛到方程 $\phi(t_p) = 0$ 的唯一解处. 由于二阶导大于 0, 为加速求解过程, 所选初值 t_p^0 最好保证 $\phi(t_p^{(0)}) > 0$, 一个合理的选择为

$$t_p^{(0)} = \frac{\|\mathbf{g}_p\|_2 - \lambda}{\gamma_1}.$$

5.5 定理 4 的证明

证明 记 $\mathbf{d}^{(m)} = \mathbf{d}_F(\boldsymbol{\beta}^{(m)}; q^{(m)})$. 假定 $\{\mathbf{d}^{(m)}\} \not\rightarrow 0$, 则存在 $\delta > 0$, $\|\mathbf{d}^{(m)}\| \geq \delta$. $\alpha_F^{(m)}$ 通过 Armijo 规则选择, 则存在 $\inf \alpha_{\text{init}}^{(m)} > 0$, 且存在常数 $\underline{m}$ 和 $\theta > 0$ 使得当 $m \geq \underline{m}$ 时, $\alpha_F^{(m)} < \theta < \alpha_{\text{init}}^{(m)}$, 并有

$$\Gamma\left(\boldsymbol{\beta}^{(m)} + \frac{\alpha^{(m)}}{\theta} \mathbf{d}^{(m)}\right) - \Gamma(\boldsymbol{\beta}^{(m)}) > \eta \frac{\alpha^{(m)}}{\theta} \Delta^{(m)},$$

其中 $0 < \sigma < 1$. 因此,

$$\begin{aligned} \sigma \Delta^{(m)} &= \sigma [\mathbf{s}^{(m)\top} \mathbf{d}^{(m)} + \mu \mathbf{d}^{(m)\top} F \mathbf{d}^{(m)} + \lambda T(\boldsymbol{\beta}^{(m)} + \mathbf{d}^{(m)}) - \lambda T(\boldsymbol{\beta}^{(m)})] \\ &< \frac{L(\boldsymbol{\beta}^{(m)} + (\alpha^{(m)}/\theta) \mathbf{d}^{(m)}) - L(\boldsymbol{\beta}^{(m)}) + \lambda T(\boldsymbol{\beta}^{(m)} + (\alpha^{(m)}/\theta) \mathbf{d}^{(m)}) - \lambda T(\boldsymbol{\beta}^{(m)})}{\alpha^{(m)}/\theta} \\ &\leq \frac{L(\boldsymbol{\beta}^{(m)} + (\alpha^{(m)}/\theta) \mathbf{d}^{(k)}) - L(\boldsymbol{\beta}^{(m)})}{\alpha^{(m)}/\theta} + \lambda T(\boldsymbol{\beta}^{(m)} + \mathbf{d}^{(m)}) - \lambda T(\boldsymbol{\beta}^{(m)}), \end{aligned}$$

其中 $0 < \alpha^{(m)}/\theta < 1$. 根据 $\Delta^{(m)}$ 的定义,

$$\begin{aligned} -(1-\sigma)\Delta^{(m)} + \gamma \mathbf{d}^{(m)\top} F \mathbf{d}^{(m)} &\leq \frac{L(\boldsymbol{\beta}^{(m)} + (\alpha^{(m)}/\theta) \mathbf{d}^{(m)}) - L(\boldsymbol{\beta}^{(m)})}{\alpha^{(m)}/\theta} - \mathbf{s}^{(m)\top} \mathbf{d}^{(m)}, \\ -\Delta^{(m)} &\geq (1-\gamma) \mathbf{d}^{(m)\top} F \mathbf{d}^{(m)} \geq (1-\gamma) \underline{\lambda} \|\mathbf{d}^{(m)}\|^2, \end{aligned} \quad (5.3)$$

其中 $\underline{\lambda}$ 是矩阵 F 的最小特征根. 因而,

$$((1-\sigma)(1-\gamma) + \gamma) \underline{\lambda} \|\mathbf{d}^{(m)}\|^2 \leq \frac{L(\boldsymbol{\beta}^{(m)} + (\alpha^{(m)}/\theta) \mathbf{d}^{(m)}) - L(\boldsymbol{\beta}^{(m)})}{\alpha^{(m)}/\theta} - \mathbf{s}^{(m)\top} \mathbf{d}^{(m)}.$$

两边同除以 $\|\mathbf{d}^{(m)}\|$, 在假定 $L(\boldsymbol{\beta})$ 为凸函数的前提下, 可以得到

$$((1-\sigma)(1-\gamma) + \gamma) \underline{\lambda} \|\mathbf{d}^{(m)}\| \leq \frac{L(\boldsymbol{\beta}^{(m)} + \hat{\alpha}^{(m)} \mathbf{d}^{(m)} / \|\mathbf{d}^{(m)}\|) - L(\boldsymbol{\beta}^{(m)})}{\hat{\alpha}^{(m)}} - \frac{\mathbf{s}^{(m)\top} \mathbf{d}^{(m)}}{\|\mathbf{d}^{(m)}\|},$$

其中 $\hat{\alpha}^{(m)} = \alpha^{(m)} \|\mathbf{d}^{(m)}\| / \theta$.

根据 (5.3), 以及 $-\alpha^{(m)} \Delta^{(m)} \geq (1-\gamma) \underline{\lambda} \alpha^{(m)} \|\mathbf{d}^{(m)}\|^2 > (1-\gamma) \underline{\lambda} \alpha^{(m)} \|\mathbf{d}^{(m)}\| \delta$, $(1-\gamma) \underline{\lambda} > 0$, 可得 $\{\alpha^{(m)} \|\mathbf{d}^{(m)}\|\} \rightarrow 0$, 即 $\{\hat{\alpha}_F^{(m)}\} \rightarrow 0$.

显然, $\{\mathbf{d}^{(m)} / \|\mathbf{d}^{(m)}\|\}$ 是有界的, 假定 $\{\mathbf{d}^{(m)} / \|\mathbf{d}^{(m)}\|\} \rightarrow \bar{\mathbf{d}}$, 而函数 $L(\cdot)$ 是连续的, 可知,

$$0 < ((1-\sigma)(1-\gamma) + \gamma) \underline{\lambda} \delta \leq \nabla L(\bar{\boldsymbol{\beta}})^{\top} \bar{\mathbf{d}} - \nabla L(\bar{\boldsymbol{\beta}})^{\top} \bar{\mathbf{d}} = 0,$$

产生矛盾. 因此, $\mathbf{d}^{(m)} \rightarrow 0$. 易得 $\lim_{m \rightarrow +\infty} \boldsymbol{\beta}_F^{(m)} = \hat{\boldsymbol{\beta}}$. □

参考文献

- 1 Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B Stat Methodol, 1996, 58: 267–288
- 2 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol, 2006, 68: 49–67
- 3 Meier L, Van De Geer S, Buhlman P. The group LASSO for logistic regression. J R Stat Soc Ser B Stat Methodol, 2008, 70: 53–71
- 4 Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol, 2005, 67: 301–320

- 5 Zou H. The adaptive LASSO and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429
- 6 Wang H, Leng C. A note on adaptive group LASSO. *Comput Statist Data Anal*, 2008, 52: 5277–5286
- 7 Osborne M R, Presnell B, Turlach B A. A new approach to variable selection in least squares problems. *IMA J Numer Anal*, 2000, 20: 389–403
- 8 Efron B, Hastie T, Johnstone I, et al. Least angle regression. *Ann Statist*, 2004, 32: 407–499
- 9 Park M Y, Hastie T. L_1 -regularization path algorithm for generalized linear models. *J R Stat Soc Ser B Stat Methodol*, 2007, 69: 659–677
- 10 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Statist Software*, 2010, 33: 1–22
- 11 Wasserman L, Roeder K. High-dimensional variable selection. *Ann Statist*, 2009, 37: 2083–2595
- 12 Tseng P, Yun S. A coordinate gradient descent method for non-smooth separable minimization. *Math Program*, 2009, 117: 387–423
- 13 Fu W J. Penalized regressions: The bridge versus the LASSO. *J Comput Graph Statist*, 1998, 7: 397–416
- 14 Fahrmeir L, Tutz G. *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2nd ed. New York: Springer-Verlag, 2001
- 15 Roth V, Fischer B. The group-LASSO for generalized linear models: Uniqueness of solutions and efficient algorithms. In: *Proceedings of the 25th International Conference on Machine Learning*. New York: Association for Computing Machinery, 2008: 848–855
- 16 Qin Z, Scheinberg K, Goldfarb D. Efficient block-coordinate descent algorithms for the group LASSO. *Math Program Comput*, 2013, 5: 143–169
- 17 Li T, Shi Y, Chen B. Homeownership and happiness: Theory and evidence from China (in Chinese). *Econ Res J*, 2011, 9: 69–82

An algorithm for the estimation of regularization paths of generalized linear models with group LASSO penalty

MA JingYi, ZHANG XinLian, SU Zhi & LIU YiWen

Abstract Computing the regularization paths of generalized linear models (GLM) with group LASSO penalty can be decomposed into two problems: Selecting the path parameter λ and computing group LASSO solution $\hat{\beta}(\lambda)$ given λ . In practice, grid method is usually used to solve the first one and coordinate descent algorithm based on the first order Taylor expansion of loss function of GLM is then used to solve the second. This paper aims at proposing algorithms that solve these two problems more efficiently. Firstly, we give a path following algorithm that attempts to find the λ 's that correspond to the change of active set. Secondly, we take advantage of the properties of GLM, and use second-order, instead of first-order, Taylor approximation of the loss function of GLM in coordinate descent method to achieve better precision in less time. Simulated and real data sets show that our algorithm is capable of efficiently pinpointing the critical λ 's that pair with changes of active set and that our proposed coordinate descent algorithm based on second-order approximation is competitive in speed compared with that based on second-order approximation.

Keywords group LASSO, generalized linear models, regularization path, coordinate descent, second-order Taylor approximation

MSC(2010) 62J07

doi: 10.1360/N012013-00101