应用概率统计
2025 年 1 月

Chinese Journal of Applied Probability and Statistics
doi: 10.12460/j.issn.1001-4268.aps.2025.2023050

基于 GPGN 算法的泊松回归稀疏优化 *

赵子榕　　王思洋†
中央财经大学统计与数学学院, 北京, 100081

摘　要: 泊松回归模型作为广义线性回归模型之一, 被广泛应用于计数型数据分析. 随着计算机技术的迅速发展, 获取和存储的变量越来越多, 所建立模型越来越复杂. 针对泊松回归模型的稀疏优化问题, 本文考虑带有 L_0 惩罚的泊松回归稀疏约束模型, 应用二阶贪婪投影梯度牛顿 (Greedy Projected Gradient Newton 简称 GPGN) 算法估计参数. 通过在合成数据集进行模拟研究说明算法的有效性, 并将泊松回归应用于基于 WIFI 信号预测楼层的建模分析, 验证了 GPGN 算法在泊松回归稀疏约束优化问题中的优良表现.

关键词: GPGN 算法; 泊松回归模型; L_0 惩罚; 稀疏约束
中图分类号: O212.4

英文引用格式: TANG Y C, ZHAO W, CAI S Y. Sparse Optimization for Poisson Regression Based on GPGN Algorithm [J]. *Chinese Journal of Applied Probability and Statistics*, 2025: 1–11. (in Chinese)

1　引言

泊松回归主要用于分析服从泊松分布的因变量与影响其取值的自变量之间变化关系的一种模型, 即分析单位时间 (或空间) 内某稀有事件的发生次数, 该模型可用于罕见疾病发病数、自然灾害发生次数、互联网的客流量等影响因素分析. 随着计算机技术的迅猛发展, 数据的搜集和存储越来越便利, 随着变量维数的增加可建立的模型也越复杂, 通常泊松回归模型假设样本量 n 大于自变量的个数 p, 因而当自变量个数 p 大于样本量 n 时需要寻找一种有效的方法来估计模型参数. 当数据的变量维数 p 大于样本量 n 时, 模型中变量越多, 模型的偏差越小, 但是过多的变量会导致模型过拟合, 降低模型的预测能力和解释性. 传统的变量筛选方法主要有信息准则和惩罚方法, 其中信息准则包含了 AIC[1]和 BIC[2]等, 惩罚方法主要有 LASSO[3], SCAD[4]和自适应 LASSO[5]等. 这些变量筛选方法用于处理变量维数 p 大于样本量 n 的回归模型需要结合 SIS[6]扫描方法降低变量维数, 这类通过两步筛选变量的方法在建模过程中不够直观.

针对泊松回归的稀疏优化问题也有一些研究, Jia 等[7]基于 L_1 惩罚提出了惩罚加权函数得分法, 证明了该方法的估计量具有相合性, 并给出了收敛速度, 可以将该方法推广到其他广义线性回归模型. Saishu 等人[8]提出了一种基于混合整数优化 (MIO) 的方法, 该方法

* 国家自然科学基金项目 (批准号: 12031016, 11971324, 11971504) 资助.
† 通讯作者, E-mail: siyangw@163.com.
本文 2023 年 5 月 11 日收到, 2023 年 7 月 9 日收到修改稿, 录用.

通过将对数似然函数进行分段线性近似, 导出了混合整数二次优化（MIQO）公式, 其中二次约束用于控制回归系数的范围, 整数变量则用于选择要保留的特征. 张露露与黄希芬[9]提出了一种新的基于 MM 算法的高维泊松回归模型, 该模型考虑了 SCAD 和 MCP 两种惩罚, 并利用组装分解技术将高维优化问题转化为低维优化问题. 该模型不仅克服了传统 Newton-Raphson 算法在矩阵求逆时的计算困难, 而且还解决了目标函数的零点奇异性问题. 已有方法提高了计算速度, 但无法直接用于估计变量维数 p 大于样本量 n 的泊松回归模型参数.

本文所提出的泊松回归稀疏优化方法, 是在损失函数中加入 L_0 惩罚, 即变量选择中常见的最优子集选择, 可直接用于估计变量维数 p 大于样本量 n 的泊松回归模型参数. 该方法筛选变量直观上更易理解, 通过限制未知参数的 L_0 范数来实现, 表达式如下:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} l(\boldsymbol{\beta}) \text{ 满足 } \|\boldsymbol{\beta}\|_0 \leqslant s$$

其中 $l(\boldsymbol{\beta})$ 是参数 $\boldsymbol{\beta}$ 的凸损失函数, $\boldsymbol{\beta}$ 是未知参数向量, $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} I\{\beta_j \neq 0\}$ 表示参数中非零元素的个数, s 为给定的正整数. 在本文中, 稀疏度 s 定义为参数向量中非零元素的个数. 泊松回归通过稀疏优化, 减少了模型参数的数量和复杂度, 在保证预测精度的同时提高了模型的可解释性.

稀疏优化属于非凸非连续优化, 是一个 NP 难问题. 早期求解带有 L_0 稀疏约束的压缩感知模型主要使用一阶贪婪算法[10–11], 随着稀疏优化的推广, 针对松弛模型设计了快速有效的算法, 如迭代阈值算法和近端梯度方法等[12]. 本文所选择的稀疏优化方法 GPGN 算法, 同时利用了损失函数的梯度向量和海森矩阵, 该算法应用于稀疏逻辑回归, 其估计具有全局收敛性、局部二次收敛性以及有限识别性[13], 这些优良性质使 GPGN 算法具有更快的迭代速度和更准确的迭代结果.

本文内容安排如下: 第一节介绍了泊松回归的变量筛选方法以及稀疏优化算法的研究现状. 第二节主要介绍了泊松回归模型的稀疏优化问题, 讨论了泊松回归的 GPGN 算法以及算法收敛性的条件. 在第三节模拟与实证中, 通过合成数据集验证了该方法的有效性, 同时将泊松回归应用于基于 WIFI 信号识别楼层的建模分析. 第四节是结论部分.

2 泊松回归与 GPGN 算法

假设 $(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)$ 是独立同分布的随机向量, $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$ 为其观测其中计数型响应变量 Y_i 取不同值的概率服从泊松分布:

$$Y_i = y_i \mid \boldsymbol{X}_i \sim \text{Poisson}\left(\lambda\left(\boldsymbol{x}_i\right)\right),\ i = 1, \cdots, n,$$

在给定样本 $\boldsymbol{x}_i$ 后, Y_i 取不同值的条件概率为:

$$\mathsf{P}\left\{Y_i = y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right\} = \frac{\lambda\left(\boldsymbol{x}_i\right)^{y_i}}{y_i!} \exp\left\{-\lambda\left(\boldsymbol{x}_i\right)\right\},\ y_i = 0, 1, 2, \cdots,$$

其中 $\log\{\lambda(\boldsymbol{x}_i)\} = x_i^\top\boldsymbol{\beta}$, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\top$ 为 p 维向量, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$ 为泊松回归模型中的待估参数. 为了估计未知参数向量 $\boldsymbol{\beta}$, 计算似然函数,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\lambda(\boldsymbol{x}_i)^{y_i}}{y_i!} \exp\{-\lambda(\boldsymbol{x}_i)\},$$

对似然函数取对数可以得到对数似然函数,

$$\log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} [y_i \log\{\lambda(\boldsymbol{x}_i)\} - \log\{y_i\} - \lambda(\boldsymbol{x}_i)], \tag{1}$$

并通过对数似然函数 (1) 求极值得到参数向量 $\boldsymbol{\beta}$ 的极大似然估计. 在实际应用中, 通过对数似然函数可定义损失函数, 泊松回归模型的损失函数为:

$$l(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i\boldsymbol{\beta}^\top\boldsymbol{x}_i - e^{\boldsymbol{\beta}^\top\boldsymbol{x}_i}\right). \tag{2}$$

记自变量观测值矩阵 $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)^\top \in \mathbb{R}^{n\times p}$ 和 $y = (y_1, y_2, \cdots, y_n)^\top$, 损失函数 (2) 经过推导可得:

引理 1　损失函数 (2) 二次连续可微并具有以下基本性质:

i. 梯度 $\nabla l(\boldsymbol{\beta})$ 为

$$\nabla l(\boldsymbol{\beta}) = \frac{\boldsymbol{X}^\top(h(\boldsymbol{\beta}) - y)}{n},$$

其中 $h(\boldsymbol{\beta})$ 是一个向量, 向量中的元素为 $h(\boldsymbol{\beta})_i = \exp\left(\boldsymbol{\beta}^\top\boldsymbol{x}_i\right), i = 1, \cdots, n.$

ii. 海森矩阵 $\nabla^2 l(\boldsymbol{\beta})$ 为

$$\nabla^2 l(\boldsymbol{\beta}) = \boldsymbol{X}^\top D(\boldsymbol{\beta})\boldsymbol{X}/n,$$

其中 $D(\boldsymbol{\beta})$ 是一个对角矩阵,

$$(D(z))_{ii} = e^{\boldsymbol{\beta}^\top\boldsymbol{x}_i}, \quad i = 1, \cdots, n.$$

对于带有 L_0 惩罚的泊松回归稀疏约束模型, 可以按如下方式定义:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} l(\boldsymbol{\beta}), \text{ 满足 } \|\boldsymbol{\beta}\|_0 \leqslant s, \tag{3}$$

GPGN 算法达到最优, 损失函数 (2) 需要满足一定的正则条件, 根据 GPGN 算法理论结果[12], 可以给出 GPGN 算法的最优性条件如引理 2 所述。

引理 2　损失函数 $l(\boldsymbol{\beta})$ 二次连续可微, 且 $l(\boldsymbol{\beta})$ 在 R^p 上是强光滑, 同时存在常数 λ_x, 使得

$$\|\nabla l(\boldsymbol{\beta}) - \nabla l(\boldsymbol{\beta}^*)\| \leqslant \lambda_x\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|.$$

在广义线性回归模型的相关文献中, 许多学者在证明参数估计的理论性质时, 对自变量观测值矩阵 $\boldsymbol{X}$ 施加不同的限制条件. 本文为了保证泊松回归稀疏约束模型 (3) 满足引

理 2, 参考陈夏和崔艳[14]的限制条件, 对泊松回归中自变量观测值向量 $\boldsymbol{X}$ 施加如下条件: 对任意 $i \in \{1,2,\cdots,n\}$, 存在 $M \in R$, 使得 $\max_{1\leqslant j\leqslant p}|x_{ij}| \leqslant M$, 因此可得:

$$\exp\{\boldsymbol{\beta}^{\top}\boldsymbol{x}_i\} \leqslant \exp\left\{\sum_{j=1}^{p}|\beta_j||x_{ij}|\right\} \leqslant \exp\left\{M\sum_{j=1}^{p}|\beta_j|\right\}.$$

为了描述方便, 将上式中常数 $\exp\{M\sum_{i=1}^{p}|\beta_j|\}$ 记为 K, 参考文献[12], 可以证明泊松回归稀疏约束模型 (3) 满足定理 3:

定理 3 模型 (3) 中的损失函数 $l(\boldsymbol{\beta})$ 是二次连续可微的, 且对任意的 $i \in \{1,2,\cdots,n\}$, 存在 $M \in \mathbb{R}$, 有 $\max_{1\leqslant j\leqslant p}|x_{ij}| \leqslant M$ 时, 损失函数具有以下性质: 损失函数 $l(\boldsymbol{\beta})$ 在 $\mathbb{R}^p$ 上是强光滑的, 记 $\lambda_x = \frac{K}{n}\lambda_{\max}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)$, 以及任意的 $\boldsymbol{\beta}, \boldsymbol{\beta}^* \in \mathbb{R}^p$

$$l(\boldsymbol{\beta}) \leqslant l(\boldsymbol{\beta}^*) + \langle\nabla l(\boldsymbol{\beta}^*), \boldsymbol{\beta}-\boldsymbol{\beta}^*\rangle + (\lambda_x/2)\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2, \tag{4}$$

并且

$$\|\nabla l(\boldsymbol{\beta}) - \nabla l(\boldsymbol{\beta}^*)\| \leqslant \lambda_x\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|.$$

证明: 对于任意的 $\boldsymbol{\beta}, \boldsymbol{\beta}^* \in \mathbb{R}^p$, 通过中值定理, 存在 $\xi \in (\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ 使得

$$\nabla l(\boldsymbol{\beta}) - \nabla l(\boldsymbol{\beta}^*) = \nabla^2 l(\xi)(\boldsymbol{\beta}-\boldsymbol{\beta}^*),$$

因为

$$\left\|\nabla^2 l(\xi)\right\| \leqslant \frac{\lambda_{\max}(D(\xi))}{n}\left\|X^{\top}X\right\| \leqslant \frac{K}{n}\lambda_{\max}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right),$$

其中, $D(\xi)$ 为引理 1 所定义的矩阵. 那么,

$$\|\nabla l(\boldsymbol{\beta}) - \nabla l(\boldsymbol{\beta}^*)\| \leqslant \left\|\nabla^2 l(\xi)\right\|\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\| \leqslant \frac{K}{n}\lambda_{\max}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\| = \lambda_x\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|,$$

在上式中, 将 $\frac{K}{n}\lambda_{\max}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)$ 记为了 λ_x, 因此, 目标函数的梯度 $\nabla l(\boldsymbol{\beta})$ 是利普西兹连续的, 其利普西兹常数为 λ_x. 基于已有文献[15] 的引理 2.3, 可以证明不等式 (4) 成立. 损失函数 $l(\boldsymbol{\beta})$ 在 $\mathbb{R}^p$ 上是强光滑的, 且参数为 λ_x. □

根据以上内容可以得出结论, 在 $\max_{1\leqslant j\leqslant p}|x_{ij}| \leqslant M$ 条件下, 稀疏约束泊松回归模型满足引理 2, 即满足 GPGN 算法的正则条件. 在实证分析中, 由于 $\lambda_x = \frac{K}{n}\lambda_{\max}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)$ 是未知的, 所以在实际计算过程中, 需要选择合适的统计量替代 λ_x 的理论值. 在证明中, 对 $\exp\left\{\boldsymbol{\beta}^{\top}\boldsymbol{x}_i\right\}$ 的取值进行限制, 根据前面的讨论, 泊松回归模型的参数 $\lambda(\boldsymbol{x}_i) = e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}$. 根据泊松分布的性质可知, $\mathsf{E}(y) = \lambda(\boldsymbol{x}_i)$, 而 $\bar{y} < \max(y)$. 本文选择了样本中响应变量 Y 的最大值 $\max\{y_i\}$ 来控制 $\exp\left\{\boldsymbol{\beta}^{\top}\boldsymbol{x}_i\right\}$.

针对该模型, 定义稀疏集 P_s, Wang 等[13]提出了贪婪投影梯度牛顿算法 (GPGN), GPGN 方法是投影梯度方法和牛顿法的结合, 具有良好的理论结果和数值模拟表现, 求解泊松回归模型的算法框架如表 1 所示:

表 1 求解泊松回归的 GPGN 算法框架

1: 初始化 $\boldsymbol{\beta}^0$, $1<\tau_0<\frac{1}{\max\{y_i\}\lambda(\boldsymbol{X}^\top\boldsymbol{X})}$, $0<\sigma\leqslant 1$, $0<\gamma<1$, $\epsilon>0$, 令 k 初值为 0;
2: (梯度步) 计算 $\tilde{\boldsymbol{\beta}}^{k+1}=P_s\left(\boldsymbol{\beta}^k-\tau_k\nabla l\left(\boldsymbol{\beta}^k\right)\right)$, 其中 $\tau_k=\tau_0\gamma^{q_k}, q_k$ 是使得下式成立的最小的整数 q, $l\left(\boldsymbol{\beta}^k\left(\tau_0\gamma^{q_k}\right)\right)\leqslant l\left(\boldsymbol{\beta}^k\right)-\frac{\sigma}{2}\left\|\boldsymbol{\beta}^k\left(\tau_0\gamma^{q_k}\right)-\beta^k\right\|^2$, 其中 $\boldsymbol{\beta}^k(\tau):=P_s\left(\boldsymbol{\beta}^k-\tau\nabla l\left(\boldsymbol{\beta}^k\right)\right)$
3: (牛顿步) 若 $\operatorname{supp}\left(\tilde{\boldsymbol{\beta}}^{k+1}\right)=\operatorname{supp}\left(\boldsymbol{\beta}^k\right)$, 那么
$\hat{\boldsymbol{\beta}}^{k+1}_{\tilde{\Gamma}^{k+1}}=\tilde{\boldsymbol{\beta}}^{k+1}_{\tilde{\Gamma}^{k+1}}-\left(\nabla^2_{\tilde{\Gamma}^{k+1}}l\left(\tilde{\boldsymbol{\beta}}^{k+1}\right)\right)^{-1}\nabla_{\tilde{\Gamma}^{k+1}}l\left(\tilde{\boldsymbol{\beta}}^{k+1}\right)$,
其中 $\tilde{\Gamma}^{k+1}$ 是 $\tilde{\boldsymbol{\beta}}^{k+1}$ 的支撑集. 否则, $\boldsymbol{\beta}^{k+1}:=\tilde{\boldsymbol{\beta}}^{k+1}$, 然后转步 5
4: (开关步) 若 $l\left(\hat{\boldsymbol{\beta}}^{k+1}\right)\leqslant l\left(\tilde{\boldsymbol{\beta}}^{k+1}\right)-\frac{\sigma}{2}\left\|\hat{\boldsymbol{\beta}}^{k+1}-\tilde{\boldsymbol{\beta}}^{k+1}\right\|^2$, 那么 $\boldsymbol{\beta}^{k+1}:=\hat{\boldsymbol{\beta}}^{k+1}$, 否则 $\boldsymbol{\beta}^{k+1}:=\tilde{\boldsymbol{\beta}}^{k+1}$
5: 如果 $\nabla_{\Gamma^{k+1}}l\left(\boldsymbol{\beta}^{k+1}\right)\leqslant\varepsilon$, 则停止. 否则, 令 $k:=k+1$ 然后转步 2

GPGN 算法需要事先给定稀疏度 s, 在本文中我们通过平衡算法运行时间以及模型拟合效果 (损失函数数值) 来实现稀疏集中非零参数个数的选择.

3 模拟与实证分析

在模拟数据集和真实数据集上, 本节将使用 GPGN 算法对稀疏约束泊松回归模型进行优化, 通过模拟试验可以验证 GPGN 算法在稀疏约束泊松回归模型优化问题中估计的准确性和计算效率. 与此同时, 本文还选择了以下几种算法进行对比: 利用 R 语言中的软件包 "lbfgs" 来实现 Broyden FletcherGoldfarb Shanno 优化算法 (下文简称 L-BFGS) 和 Orthant Wise Quasi Newton Limited Memory 优化算法 (下文简称 OWL-QN)[7]、GLMnet 方法[16]、Picasso 方法[17](惩罚项本文分别选择了 L_1 惩罚和 MCP 惩罚).

3.1 模拟分析

设定 $\varepsilon=10^{-6}$, 并且令初始值 $\boldsymbol{\beta}^0$ 为零. 对于稀疏泊松回归模型, 本文数据模拟参考了文献[8], 将样本量记为 n, 自变量个数记为 p, 未知参数向量的稀疏度记为 s. 假设自变量 $\boldsymbol{x}_i\sim N(0,\Sigma)$, $i=1,\cdots,n$, 其协方差矩阵为 $\Sigma=(\sigma_{ij})_{p\times p}$, 其中 $\sigma_{ij}=\rho^{|i-j|}$. 自变量间的相关性由 ρ 的取值决定, 本文考虑 $\rho\in\{0,0.35,0.7\}$ 三种情况. 本文考虑到了随机误差项的存在, 对于响应变量 Y, 参考文献利用舍入的方法生成计数变量 $y_i\in\{0\}\cup[10]$, 通过对

$$\exp\left\{(\boldsymbol{\beta}^{*\top}\boldsymbol{x}_i)/\sqrt{\boldsymbol{\beta}^{*\top}\Sigma\boldsymbol{\beta}^*}+\varepsilon_i\right\},$$

取整数 (四舍五入) 作为 y_i 的值, 其中 $\varepsilon_i\sim N(0,\sigma^2)$, σ 取 0.1. 在模拟数据集中, 为了比较不同算法的效果, 本文选取了损失函数值、运算时间、稀疏度以及 FNR、FPR[18]作为评价指标, 其中 FNR、FPR 定义如下:

$$\mathrm{FNR}=\{j:\beta_j\neq 0\text{ 但 }\hat{\beta}_j=0\},$$

$$\mathrm{FPR} = \{j : \beta_j = 0 \text{ 但 } \hat{\beta}_j \neq 0\},$$

特征之间的相关性 ρ 分别取 0, 0.35, 0.7. 在这三种相关性下, 固定样本数量 $n = 2000$ 和变量维数 $p = 6000$, 稀疏度 s 设定为 500. 与其他算法不同, GPGN 算法的使用过程中未知系数向量的稀疏度 s 需要作为超参数事先给定, 为了排除主观因素的影响, 本文选择综合比较算法运行时间和损失函数值在不同稀疏度设定下的变化趋势, 以此为依据来选择合适的稀疏度 s. 在本例中, 真实系数向量的稀疏度 s 已知, 如果假设 s 未知, 可以验证所提出稀疏度选择方法的有效性. 当 ρ 取 0 时, 随着稀疏度 s 的变化, GPGN 算法运行时间以及损失函数的取值变化情况如图 1-2.

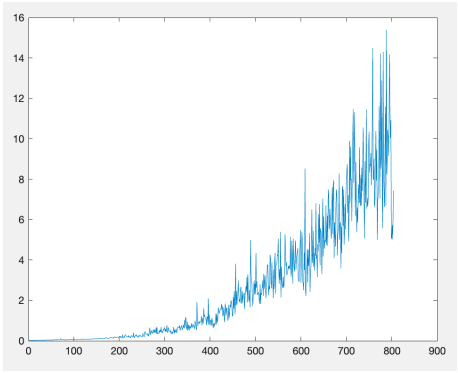

图 1 稀疏度与运算时间折线图

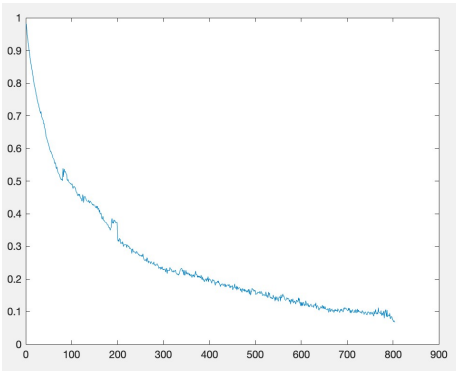

图 2 稀疏度与损失函数折线图

图 1 与图 2 的横坐标均为稀疏度, 不同的是, 图 1 的纵轴表示运算时间, 图 2 的纵轴表示损失函数. 损失函数值减少的速度随着稀疏度 s 的提高逐渐变慢, 而运算时间在稀疏度达到 400 到 500 之间后增速逐渐加快, 结合具体数值, 选定稀疏度为 439.

表 2 实验结果 $\rho = 0$

方法	损失函数值	运算时间(秒)	稀疏度	FNR	FPR
GPGN	0.170	0.998	439	0.051	0.019
L-BFGS	-0.017*	14.748	6000	0.000	0.460
OWL-QN	0.016	46.620	2518	0.019	0.180
GLMnet	0.015	1.671	1856	0.014	0.123
Picasso-L1	0.036	1.563	1274	0.012	0.073
Picasso-MCP	0.211	5.999	957	0.022	0.051

* 损失函数推导中省略了与参数无关的项, 进而计算结果出现负值, 下同.

各算法运算结果汇总于表 2, 通过结果分析得到 GPGN 算法稀疏度较小运行时间较短, 但是会把一些非零系数选择成零, 即 FNR 较高; 但是 FPR 较低, 即剔除系数为零的非重要变量效果更好, 因而模型稀疏度较小, 损失函数值较大仅低于 Picasso-MCP.

当 ρ 取 0.35 时, 随着稀疏度 s 的变化, GPGN 算法运行时间以及损失函数的取值变化情况如图 3-4.

运行时间和损失函数随着稀疏度的变化趋势与 $\rho = 0$ 时的情况类似, 在此不再赘述,

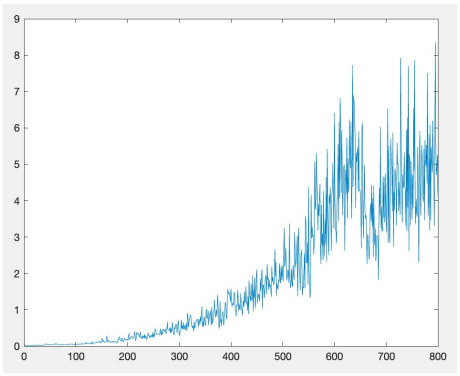

图 3 稀疏度与运算时间折线图

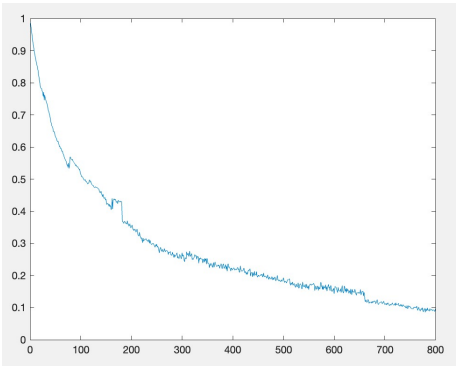

图 4 稀疏度与损失函数折线图

根据数据结果, 选定稀疏度 $s = 461$.

表 3 实验结果 $\rho = 0.35$

方法	损失函数值	运算时间(秒)	稀疏度	FNR	FPR
GPGN	0.208	1.160	461	0.050	0.027
L-BFGS	-0.102	14.597	6000	0.000	0.483
OWL-QN	-0.069	46.337	2599	0.018	0.206
GLMnet	0.061	1.705	1803	0.015	0.134
Picasso-L1	0.059	1.444	1192	0.015	0.078
Picasso-MCP	0.045	5.970	908	0.022	0.052

各算法运算结果汇总于表 3, 通过结果分析得到 GPGN 算法稀疏度较小运行时间较短, 但是会把一些非零系数选择成零, 即 FNR 较高; 但是 FPR 较低, 即剔除系数为零的非重要变量效果更好, 模型稀疏度较小, 损失函数值较大.

当 ρ 取 0.7 时, 随着稀疏度 s 的变化, GPGN 算法运行时间以及损失函数的取值变化情况如图 5-6.

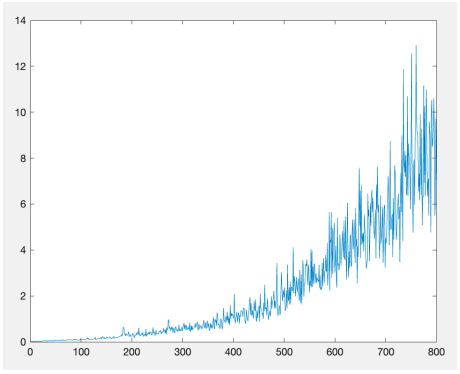

图 5 稀疏度与运算时间折线图

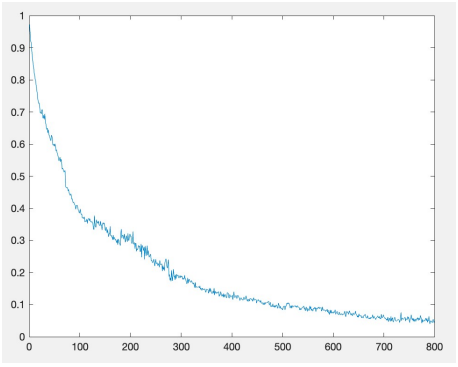

图 6 稀疏度与损失函数折线图

与前文类似, 根据数据结果, 当 $\rho = 0.7$ 时, 选定稀疏度 $s = 500$. 各算法运算结果汇总于表 4, 通过结果分析得到 GPGN 算法稀疏度较小运行时间较短, 但是会把一些非零系数

选择成零, 即 FNR 较高; 但是 FPR 较低, 即剔除系数为零的非重要变量效果更好, 模型稀疏度较小, 损失函数值较大, 仅低于 OWL-QN 和 Picasso-L1.

表 4 实验结果 $\rho = 0.7$

方法	损失函数值	运算时间(秒)	稀疏度	FNR	FPR
GPGN	0.083	1.226	500	0.047	0.038
L-BFGS	0.063	25.602	6000	0.000	0.513
OWL-QN	0.097	42.259	2951	0.020	0.266
GLMnet	0.044	1.513	1711	0.018	0.147
Picasso-L1	0.084	1.275	1052	0.018	0.092
Picasso-MCP	0.073	5.261	693	0.034	0.050

注记 1 由于 GPGN 算法需要事先给定稀疏度 s, 本文在模拟与实证分析中, 通过平衡运行时间和模型拟合效果两方面 (损失函数值) 选择的稀疏度较小, 因而 GPGN 算法的 FPR 较低, 运行时间与其他算法相比有较明显优势. 较小的稀疏度使得 GPGN 算法筛选变量时会将部分非零系数选成零, 因而相较于其他方法 FNR 偏高, 同时损失函数值略高于部分算法. 通过模拟结果我们可以看到 GPGN 算法平衡了运行效率和模型拟合效果两方面, 在后续研究中继续讨论稀疏度 s 选择方法在提高效率的同时提升模型拟合效果.

3.2 实证分析

许多应用程序需要获取用户所在的具体位置以更好地向用户提供服务, 因而用户自动定位一直是近年来的研究热点. 用户自动定位包括使用电子设备 (通常是手机) 估计用户的位置 (纬度、经度和高度). 由于移动设备中包含 GPS 传感器, 室外定位比较准确. 然而, 室内定位仍然具有一定难度, 主要是由于室内环境中 GPS 信号丢失. 尽管有一些室内定位技术和方法, 数据集主要关注基于 WLAN 指纹的技术和方法 (也称为 WiFi 指纹识别). 为了利用模型预测楼层, 本文对 UJIIndoorLoc 数据集进行建模分析, 该数据集有 19937 个样本, 528 个变量, 数据中各变量具体含义说明如表 5 所示.

表 5 数据各变量具体含义说明

变量名	变量含义
WAP001-WAP520	用户的 WLAN 信号值
经度	用户所处位置的经度值
纬度	用户所处位置的纬度值
楼层	用户在建筑内所处在的具体楼层
建筑编号	用户所处于的建筑识别编号
空间编号	用户所处于的具体空间区域（教室、办公室等）
相对位置	用户的相对位置, 1=“室内”, 2=“室外”
用户 ID	识别用户的 ID
设备 ID	用户所使用的手机设备型号
时间	数据获取时的具体时间

为了更加直观地展示楼层变量的取值分布, 绘制楼层变量的直方图见图 7, 其中横坐标

表示楼层, 纵坐标表示频数. 楼层均值为 1.675, 而从图 7 可以直观的看出, 大部分用户所处楼层为 1 层或 2 层, 并且处于 4 层的用户数量较少, 可以认为基本符合泊松分布的分布规律, 因此, 使用泊松回归模型是有意义的. GPGN 算法的稀疏度 s 需要在算法实施前作

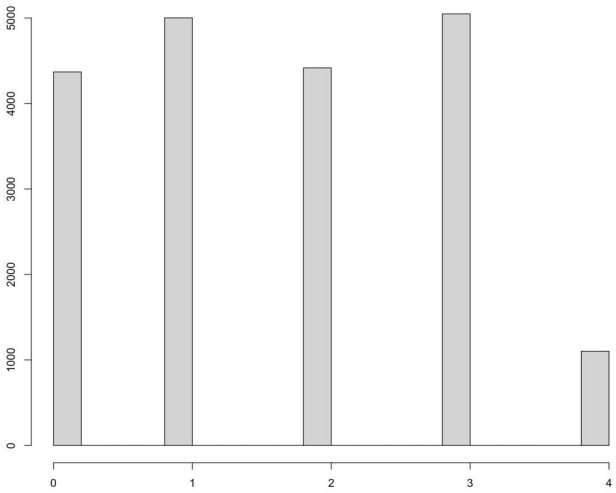

图 7 楼层直方图

为超参数给定, 对于真实数据集而言, 合适的稀疏度选取尤为重要. 与前文模拟数据集部分所述方法一致: 使用不同的稀疏度 s 进行迭代, 观察不同稀疏度下 GPGN 算法的运行结果, 比较运行时间以及最终损失函数的取值, 以此为依据, 为数据选取出合适的稀疏度.

通过实验研究, 可以将稀疏度 s 与运行时间、损失函数值之间的关系绘制为图 8-9. 图 8 与图 9 的横轴均为稀疏度 s, 纵轴分别是运算时间和损失函数值, 从这两幅图中, 可以看出随着稀疏度的提高, 运算时间波动上升, 而损失函数值在稀疏度达到 170 之后逐渐趋于一个稳定的取值, 因此, 结合以上两幅图所展示的结果, 本文选择在运算过程中将该数据集的稀疏度设定为 170.

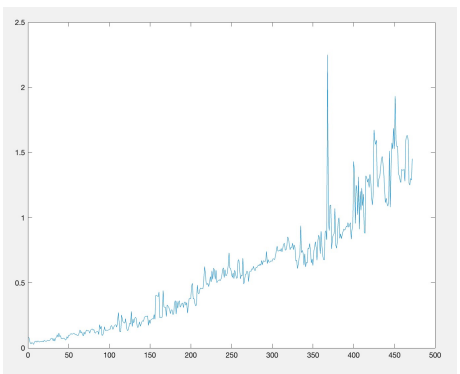

图 8 稀疏度与运算时间折线图

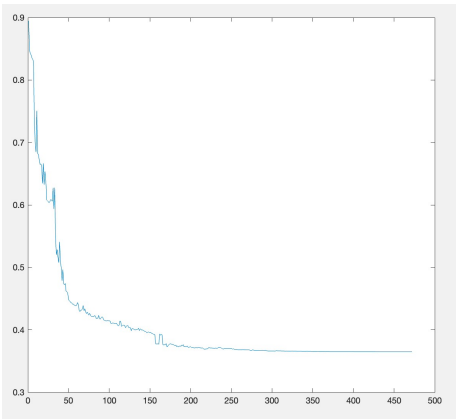

图 9 稀疏度与损失函数折线图

该数据库涵盖了三栋建筑的数据, 其中有 4 层或更多层, 本文选择将所处楼层作为响应变量, 其余经度纬度以及 WLAN 指纹解锁信号等数据均作为自变量, 设定稀疏度 $s = 170$, 所选取的比较算法为 GPGN 算法, L-BFGS 算法, OWL-QN 算法, GLMnet, Picasso-L_1, Picasso-MCP 用来进行比较的指标为损失函数值, 运算时间与稀疏度.

表 6 UJIIndoorLoc 数据集实验结果

比较算法	损失函数值	运算时间 (秒)	稀疏度
GPGN	0.372	0.242	170
L-BFGS	0.395	35.392	472
OWL-QN	0.385	50.141	469
GLMnet	0.380	0.773	462
Picasso-L_1	0.405	1.127	155
Picasso-MCP	0.391	2.550	98

从表 6 可以看出, GPGN 算法依然在运算时间上有很大优势, 并且在所有的比较算法的迭代结果中, GPGN 算法的损失函数值较小, 说明 GPGN 算法的迭代结果更加精确.

4 结论

本文考虑带有 L_0 惩罚的泊松回归稀疏约束模型, 将二阶算法 GPGN 算法应用于参数估计. 验证了在一定条件下, 泊松分布满足 GPGN 算法的最优性条件. 在模拟数据集和真实数据集上分别建立泊松回归模型, 比较 GPGN 算法与其他常用算法在泊松回归稀疏优化问题上的结果. 通过模拟和实例分析发现, 只要给定的稀疏度接近真实值, GPGN 算法相较于其他算法在运算速度与变量筛选方面有一定的优势. 当给定的稀疏度较小时, GPGN 算法的运算时间较短, FPR 相较于其他算法较低, 说明将系数为零的变量剔除至模型以外的能力相较于其他算法更强一些, 但是 FNR 与其他方法相比稍有劣势. 随着变量间相关性的提高, 在各个方法中, 各项指标的变化并不明显, 说明这些方法的运算结果受变量间相关性影响较小.

参考文献

[1] AKAIKE H. A new look at the statistical model identification [J]. *Automatic Control, IEEE Transactions on*, 1974, **19**(**6**): 716–723.

[2] SCHWARZ G. Estimating the Dimension of a Model [J]. *The Annals of Statistics*, 1978, **6**(**2**): 461–464.

[3] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. *Journal of the Royal Statistical Society, Series B*, 1996, **58**(**1**): 267–288.

[4] LI R Z, FAN J Q. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties [J]. *Publications of the American Statistical Association*, 2001, **96**(**456**): 1348–1360.

[5] ZOU H. The Adaptive Lasso and Its Oracle Properties [J]. *Publications of the American Statistical Association*, 2006, **101**(**476**): 1418–1429.

[6] FAN J Q, LV J C. Sure independence screening for ultra-high dimensional feature space (with discussion) [J]. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 2008, **70**(**5**):

849–911.

[7] JIA J Z, XIE F, XU L H. Sparse Poisson Regression with Penalized Weighted Score Function [J]. *Electronic Journal of Statistics*, 2017, **13**(**2**): 2898–2920.

[8] SAISHU H, KUDO K, TAKANO Y. Sparse Poisson regression via mixed-integer optimization [J/OL]. *PLOS ONE*, 2021, **16**. https://doi.org/10.1371/journal.pone.0249916.

[9] 张露露, 黄希芬. 基于 MM 算法的高维泊松回归模型变量选择 [J]. 统计与决策, 2021(**21**): 24–28.

[10] 于春梅. 稀疏优化算法综述 [J]. 计算机工程与应用, 2014, **50**(**11**): 210–217.

[11] 文再文, 印卧涛, 刘歆, 等. 压缩感知和稀疏优化简介 [J]. 运筹学学报, 2012, **16**(**3**): 49–64.

[12] 王锐. 稀疏逻辑回归二阶方法研究 [D]. 北京交通大学, 2021.

[13] WANG R, XIU N H, ZHANG C. Greedy Projected Gradient-Newton Method for Sparse Logistic Regression [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(**2**): 527–538.

[14] 陈夏, 崔艳. 高维广义线性模型的拟似然自适应 Lasso 估计 [J]. 陕西师范大学学报: 自然科学版, 2019, **47**(**2**): 1–9.

[15] BECK A, ELDAR Y C. Sparsity Constrained Nonlinear Optimization: Optimality Conditions and Algorithms [J]. *SIAM Journal on Optimization*, 2012, **23**(**3**): 1480–1509.

[16] BAHMANI S, RAJ B, BOUFOUNOS P. Greedy Sparsity-Constrained Optimization [J]. *Journal of Machine Learning Research*, 2013, **14**(**1**): 807–841.

[17] GE J, LI X G, JIANG H M, et al. Picasso: A Sparse Learning Library for High Dimensional Data Analysis in R and Python [J]. *Journal of Machine Learning Research*, 2019, **20**(**44**): 1–5.

[18] CHOOSAWAT C, REANGSEPHET O, SRISURADETCHAI P, et al. Performance Comparison of Penalized Regression Methods in Poisson Regression under High-Dimensional Sparse Data with Multicollinearity [J]. *Thailand Statistician*, 2020, **18**(**3**): 306–318.

Sparse Optimization for Poisson Regression Based on GPGN Algorithm

ZHAO Zirong WANG Siyang

School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 100081, China

Abstract: Poisson regression model, as one of the generalized linear regression models, is widely used in counting data analysis. With the rapid development of computer technology, more and more variables are obtained and stored, resulting in increasingly complex models. In this paper, we consider the sparsity constrained Poisson regression model with L_0 penalty, and apply the Greedy Projected Gradient Newton(GPGN) algorithm to estimate the parameters. The effectiveness of the algorithm is demonstrated through simulation research on the synthetic dataset, and Poisson regression is applied to the modeling analysis of the prediction floors based on WIFI signals. This verify the GPGN algorithm performs well in Poisson regression sparsity constrained optimization.

Keywords: GPGN algorithm; Poisson regression model; L_0 penalty; sparsity constrained

2020 Mathematics Subject Classification: 62J12