数学学报(中文版)
Acta Mathematica Sinica(Chinese Series)
ISSN 0583-1431,CN 11-2038/O1

《数学学报(中文版)》网络首发论文

题目： ODS 抽样下高维数据广义线性回归的自适应矩估计算法
作者： 朱京宇，丁洁丽
网络首发日期： 2024-09-09
引用格式： 朱京宇，丁洁丽．ODS 抽样下高维数据广义线性回归的自适应矩估计算法[J/OL]．数学学报(中文版).
https://link.cnki.net/urlid/11.2038.O1.20240906.0947.002

数 学 学 报 中文版
ACTA MATHEMATICA SINICA, CHINESE SERIES

文献标识码：A

ODS抽样下高维数据广义线性回归的自适应矩估计算法

朱京宇

武汉大学数学与统计学院 武汉 430072
广州外国语学校 广州 511455
E-mail: jyzhu814@whu.edu.cn

丁洁丽

武汉大学数学与统计学院 武汉 430072
E-mail: jlding.math@whu.edu.cn

摘 要 基于因变量抽样设计(Outcome-Dependent Sampling Design, 简称ODS抽样设计)是一种回溯性的有偏抽样方法. 对于大规模数据的研究, ODS抽样机制能够节约研究成本和提高效率. 本文探讨如何应用广义线性模型来拟合采用ODS抽样设计获取的高维数据. 受梯度下降思想的启发, 本文发展了两种改进的自适应矩估计算法来解决高维ODS数据的广义线性回归中估计的数值计算问题, 并证明了所提出算法的收敛性. 所提出的这些自适应矩估计算法避免了计算高维矩阵及其逆矩阵, 表现优良. 本文通过一系列的模拟研究展示了所提出算法的性能, 并应用所提出的算法分析了一个实际数据.

关键词 ODS抽样设计; 高维数据; 自适应矩估计算法; 广义线性模型
MR(2020)主题分类 62D05; 62J12
中图分类 O212.2

Adaptive Moment Estimation Algorithms under Outcome-Dependent Sampling Design with High-dimensional Data in Generalized Linear Regression

Jing Yu ZHU

School of Mathematics and Statistics, Wuhan University, Wuhan, 430072
Guangzhou Foreign Language School, Guangzhou, 511455
E-mail: jyzhu814@whu.edu.cn

Jie Li DING

School of Mathematics and Statistics, Wuhan University, Wuhan, 430072
E-mail: jlding.math@whu.edu.cn

通讯作者: 丁洁丽

Abstract An outcome dependent sampling (ODS) design is a biased-sampling scheme, which can save the cost and improve the efficiency in studies on large-scale data. We study how to fit the generalized linear models to high-dimensional data collected via ODS design. Inspirited the idea of gradient descent algorithm, we develop two improved adaptive moment estimation algorithms for the computation of the estimator in generalized linear regression with high-dimensional ODS data, and establish the theoretical properties. The proposed algorithms obviates the computation of some high-dimensional matrices and their inverses. We conduct simulation studies and analyze a real data example to illustrate the performance of the proposed algorithms.

Keywords outcome-dependent sampling design; high-dimensional data; adaptive moment estimation algorithm; generalized linear models.
MR(2020) Subject Classification 62D05; 62J12
Chinese Library Classification O212.2

1 引言

简单随机抽样由于其原理简单和理论成熟等优点, 得到了非常广泛的应用. 然而, 在很多基于大规模数据的队列研究领域, 使用简单随机抽样会有明显的局限性. 例如, 在环境学、经济学以及流行病学等领域的大型队列研究中, 研究人员们往往需要探究响应变量与一些影响因素之间的关联关系, 其中某些关键的影响因素或者协变量的调查成本相对高昂 (比如: 污染物含量、供给量、生物位点等). 若使用传统的简单随机抽样往往会导致研究过于昂贵而超出预算. 这种情况下, 研究人员更倾向于选用一些有偏抽样设计以提高实验效率并节约成本. ODS 抽样设计是一种回溯性的有偏抽样方法, 其中测量昂贵协变量的概率依赖于其因变量的观测值. 很多研究中, 因变量的采集往往比较便宜, 因此研究人员可以观测每个研究对象的因变量. ODS 抽样机制为: 首先, 在全队列中抽取一个简单随机样本(SRS样本)来提供全局信息; 之后, 将队列中的个体按照其因变量的观测值分成若干层, 从每层中抽取补充样本; 最后, 昂贵协变量的采集仅对 SRS 样本和补充样本中的个体进行. 这种 ODS 有偏抽样可以将研究资源集中在包含有更大信息量的群体上, 从而达到节约成本和提高效率的目的. 目前关于 ODS 抽样设计已经有了很多研究成果 (Zhou等人, 2002[22]; Brelsow等人, 2003[1]; Song等人, 2009[16]; Qin&Zhou, 2010[14]; Tan 等人, 2016[17]; Ding等人, 2017[4]; Yan等人, 2017[21]; Cai等人, 2019[3]; Sauer等人, 2022[15]; 等等). 这些已有研究均表明, ODS 抽样是具有成本效益的有偏抽样方法.

在大维或高维数据的大型队列研究中, 本文首先考虑采用 ODS 抽样来获取数据并应用广义线性模型来拟合这些 ODS 数据. 对于回归参数的统计推断, 我们采用 Yan 等人(2017)[21] 提出的一种半参数经验似然估计方法. 当模型参数的统计推断方法建立之后, 一个实际中迫切需要解决的问题是估计的数值计算问题. 在数据维度较低时, Newton-Raphson 算法是应用最为广泛的数值算法之一. 但是, 当数据维数较高时, Newton-Raphson 算法很容易遇到一些问题. 例如: 计算中一些大维或高维矩阵计算不可逆; 算法的收敛性过多依赖初值的选取, 等等. 近年来, 很多数据降维的方法被提出和研究 (Tibshirani, 1996[18], 1997[19]; Fan&Li, 2001[7], 2002[8]; Fan&Lv, 2008[9]; 等等). 然而, 这些降维方法大多要求稀疏性假设. 当稀疏

性假设不成立或者降维之后数据维数仍然较大时, 我们需要建立新的算法来实现感兴趣估计的数值计算.

梯度下降法作为一种最经典的优化算法得到了广泛的应用. 顾名思义, 梯度下降法是一种通过沿着目标函数的负梯度方向前进搜索极小值的算法, 需要走多远由其学习率或步长决定 (Lange, 2004[11]; Boyd&Vandenberghe, 2004[2]). 然而, 传统梯度下降算法的步长是固定的. 如果步长设定过大, 在计算中可能会在最优解附近震荡而无法收敛; 反之, 如果步长设定过小, 则可能会导致算法收敛速度过慢, 有时可能会停滞在局部最小值处. 近年来, 一系列对梯度下降法进行改进的算法和相关理论得到了蓬勃发展 (Nesterov, 1983[12]; Qian, 1999[13]; Duchi等人, 2011[6]; Tieleman&Hinton, 2012[20]; Kingma&Ba, 2015[10]; Dozat, 2016[5]). 这些算法通过修正搜索方向、学习率的动态选取等等来加速算法的收敛.

在高维ODS数据的统计分析中, 由于ODS抽样中有偏抽样机制的存在和ODS 样本外关键协变量信息的缺失, 导致我们感兴趣的估计的数值计算更容易遭遇上述提到的各种问题. 在本文中, 我们发展两种改进的梯度算法来解决高维ODS数据的广义线性回归中估计的数值计算问题. 自适应矩估计算法 (Adaptive Moment Estimation Algorithm, 以下简写为Adam算法)(Kingma&Ba, 2015[10])是梯度下降法的改进算法中很为高效的一种. 它将搜索方向和学习率结合在一起考虑, 结合动量梯度下降法使用梯度的指数加权并使用自适应学习率的调整. Nadam算法 (Nesterov-Accelerated Adaptive Moment Estimation Algorithm, 以下简写为Nadam算法) (Dozat, 2016[5])是Adam算法的一种变体. 它在Adam算法的基础上融合了Nesterov动量项 (Nesterov, 1983[12]), 在迭代中通过考虑未来梯度的因素来加速Adam算法的收敛. 受这两种改进梯度下降思想的启发, 在本文中, 对于高维ODS数据的广义线性回归中目标函数的极大值问题, 我们给出了计算其极大半参数经验似然估计的Adam算法和Nadam算法. 所提出的这些算法避免了高维矩阵及其逆矩阵的计算, 抑制了算法在极值处的震荡, 算法表现良好且具有较好的收敛速度.

本文其余部分结构如下. 在第二节中, 我们介绍ODS抽样设计和阐述广义线性模型下参数的估计方法. 在第三节中, 建立高维ODS数据的Adam算法和Nadam算法, 并获得所提出算法的收敛性. 在第四节中, 通过模拟研究展示所提出的算法的实际表现. 在第五节中, 应用所提出的方法和算法分析一个实际数据. 在第六节中, 对本文工作做出总结和讨论. 定理的证明总结在附录中.

2 抽样与估计方法

ODS抽样是一种回溯性的两阶段有偏抽样机制. 在第一阶段，我们首先测量队列中所有个体的因变量的值; 在第二阶段, 我们根据第一阶段测量所得的因变量信息抽取部分个体, 然后观测其协变量的值. 这种ODS抽样机制通过允许每个个体的入样概率依赖于因变量的信息来提高研究的效率. 当大型队列研究中对昂贵协变量的采集产生大量费用时, ODS抽样可以在认为包含更多信息的群体中抽取更多的个体，从而节省研究成本.

具体而言, 假设研究队列包含有N个个体, 对于第i个个体, 记其因变量为Y_i, 协变量为X_i $(i = 1, \cdots, N)$. 事先选取常数$\{a_0, \cdots, a_K\}$, 满足$-\infty = a_0 < \cdots < a_{k-1} < a_k < \cdots < a_K = +\infty$. 基于这些分割点将因变量的值域分为$K$个互斥且完备的区间, $A_k = (a_{k-1}, a_k]$, $k =$

$1,\cdots,K$. 首先, 测量队列中每个个体的因变量Y_i, $i=1,\cdots,N$. 根据这些因变量的观测值, 我们将全队列划分为K层. 然后, 从全队列中应用简单随机抽样来选取一个样本量为n_0的样本, 记为SRS样本. 接着, 分别在第k层中抽取样本量为n_k的样本作为补充样本. 样本容量n_0, n_k $(k=1,\cdots,K)$均为事先给定. SRS样本与每层中的补充样本组成ODS样本. 最后, 仅对ODS样本中的个体测量其协变量的值. 因此, ODS抽样设计的样本观测数据结构如下:

$$\text{ODS 样本}\ \begin{cases}\text{SRS 样本:}\quad (Y_i,X_i),\ i\in S_0;\\ \text{补充样本:}\quad (Y_i,X_i|Y_i\in A_k),\ i\in S_k,\ k=1,\cdots,K,\end{cases}\tag{2.1}$$

其中S_0和S_k分别表示SRS样本和第k层补充样本的指标集. 记$n_V=\Sigma_{k=0}^{K}n_k$为ODS样本的样本量.

我们考虑如下广义线性模型, 给定X_i的情况下, 因变量Y_i的条件分布密度有如下形式:

$$f_\beta(Y_i|X_i)=\exp\{g(X_i'\beta)Y_i-b(g(X_i'\beta))\},\tag{2.2}$$

其中函数$b(\cdot)$形式已知, β为p维待估参数. 函数$g(\cdot)$为联系函数, 它构建了因变量Y_i与协变量X_i和待估参数β之间的联系. 在广义线性模型框架下, 有

$$\mathrm{E}(Y_i|X_i)=\dot{b}(g(X_i'\beta)),\quad \mathrm{Var}(Y_i|X_i)=\ddot{b}(g(X_i'\beta)),$$

其中$\dot{b}(\cdot)$与$\ddot{b}(\cdot)$分别表示函数$b(\cdot)$的一阶导数和二阶导数. 这说明广义线性模型是被广泛应用的线性模型的推广.

在广义线性模型(2.2)的框架下, 基于ODS样本的观测数据(2.1) 的似然函数有如下形式:

$$L(\beta,q)=\prod_{i\in S_0}f_\beta(Y_i|X_i)q(X_i)\times\prod_{k=1}^{K}\prod_{i\in S_k}\frac{f_\beta(Y_i|X_i)q(X_i)}{P(Y_i\in A_k)},$$

其中$q(x)$表示协变量X的边际分布密度函数. 经过一些推导和计算, 我们可得相应的对数似然函数如下:

$$\ell(\beta,Q)=\sum_{k=0}^{K}\sum_{i\in S_k}\{\ln f_\beta(Y_i|X_i)+\ln q(X_i)\}-\sum_{k=1}^{K}n_k\ln\left\{\int_{\chi}\int_{A_k}f_\beta(y|x)dydQ(x)\right\},\tag{2.3}$$

其中, $Q(x)$表示协变量X的边际累积分布函数. 假设$q(x)$和$Q(x)$与参数β无关. 注意到, 函数$Q(x)$无法从似然函数(2.3)中分离出来, 故上述似然函数$\ell(\beta,Q)$无法直接对β极大化来实现参数的估计.

在广义线性模型下, Yan等人(2017)[21]为了解决ODS抽样中参数的估计问题, 提出了一种半参数经验似然法. 首先, 令$p_i=q(X_i)=dQ(X_i)$, $i\in S_k$, $k=0,\cdots,K$, 代入对数似然函数(2.3), 得到

$$\ell(\beta,\{p_i\})=\sum_{k=0}^{K}\sum_{i\in S_k}\{\ln f_\beta(Y_i|X_i)+\ln p_i\}-\sum_{k=1}^{K}n_k\ln\left\{\sum_{k=0}^{K}\sum_{i\in S_k}p_iP_k(X_i,\beta)\right\},\tag{2.4}$$

其中$P_k(x,\beta)=\int_{A_k}f_\beta(y|x)dy$, 且需满足约束$\left\{p_i\geq 0,\ i\in S_k,\ k=0,\cdots,K;\ \sum_{k=0}^{K}\sum_{i\in S_k}p_i=1\right\}$. 然后, 使用拉格朗日乘子法来解决这个带约束的极值问题, 应用Yan等人(2017)[21]中的具体推导过程, 我们可得到如下目标似然函数:

$$\ell(\beta,\pi)=\sum_{k=0}^{K}\sum_{i\in S_k}\ln f_\beta(Y_i|X_i)-\sum_{k=0}^{K}\sum_{i\in S_k}\ln\left[n_0\sum_{k=1}^{K}\left\{1+\frac{n_k}{n_0\pi_k}\right\}P_k(X_i,\beta)\right]-\sum_{k=1}^{K}n_k\ln\pi_k,\tag{2.5}$$

其中冗余参数 $\pi_k = \int_{\chi} P_k(x,\beta)dQ(x)$. 将得到的目标似然函数(2.5)关于参数 $\eta = (\beta', \pi')'$ 极大化, 得到 η 的极大半参数经验似然估计, 记为 $\hat{\eta}_V = (\hat{\beta}_V', \hat{\pi}_V')'$, 其中 $\hat{\beta}_V$ 即为相应的回归参数 β 的极大半参数经验似然估计.

Yan等人(2017)[21]证明了 $\hat{\eta}_V$ 的渐近性质. 用上标 "0" 代表参数的真实值, 如 η^0 表示参数 η 的真值. 在一些正则条件下, $\hat{\eta}_V$ 依概率收敛到 η^0, 且

$$\sqrt{n_V}(\hat{\eta}_V - \eta^0) \xrightarrow{d} N(0, \Omega(\eta^0)),$$

其中渐近方差 $\Omega(\eta^0) = \Psi^{-1}(\eta^0)\{\Sigma_1(\eta^0) + \Sigma_2(\eta^0)\}\Psi^{-1}(\eta^0)$, $\Psi(\eta^0)$, $\Sigma_1(\eta^0)$ 和 $\Sigma_2(\eta^0)$ 的具体定义请见附录.

实际应用中, $\hat{\eta}_V$ 的计算需要采用数值计算方法, 最为广泛使用的是 Newton-Raphson (N-R) 算法. 然后，在大维或者高维数据情形下, N-R 算法可能会遭遇一些问题, 因此, 我们将建立一种新的数值算法来解决 ODS 抽样下参数的极大半参数经验似然估计的数值计算问题.

3 数值算法

为了实现 ODS 抽样设计下的参数推断, 我们需要解决如下优化问题:

$$\hat{\eta}_V = \arg\max_{\eta} \ell(\eta), \tag{3.1}$$

其中, 目标似然函数 $\ell(\eta)$ 如式(2.5)所示且关于 η 是上凸函数 (Yan等人, 2017[21]). 由于相应的得分方程 $\nabla\ell(\eta) = 0$ 没有显式解, 我们需要应用数值方法计算 $\hat{\eta}_V$ 的估计值. 最为广泛使用的是 N-R 算法, 其迭代更新公式如下:

$$\eta_{t+1} = \eta_t + [H(\eta_t)]^{-1}\nabla\ell(\eta_t),$$

其中 η_t 表示待估参数 η 的第 t 次迭代结果, $H(\eta) = -\nabla^2\ell(\eta)$ 是观测信息矩阵. 然而, 如果数据的维数较高, 信息矩阵 $H(\eta)$ 有可能出现计算不可逆的情况. 虽然存在一些针对高维数据降维的方法, 但这些现有的方法往往需要稀疏性假设. 在实际应用中, 数据集可能不满足稀疏性假设或者在进行数据降维之后, 数据集的维数可能仍然很大.

近年来, 针对高维数据数值计算问题, 梯度下降算法发展迅速. 受传统梯度下降思想的启发，我们首先提出如下更新公式来解决我们的目标优化问题(3.1):

$$\eta_{t+1} = \eta_t + \alpha\nabla\ell(\eta_t), \tag{3.2}$$

其中 α 是事先设定的步长或学习率. 可以看出, 我们提出的这种梯度"上升"法就是沿着目标函数 $\ell(\eta_t)$ 的正梯度方向搜索极大值. 这种算法避免了 N-R 算法中计算信息矩阵 $H(\eta)$ 以及其逆矩阵的问题. 然后, 这种梯度算法使用了固定步长, 这往往会导致算法的收敛速度非常缓慢. 因此, 我们进一步考虑改进的梯度下降算法来解决我们的目标优化问题.

自适应矩估计 (Adam) 算法是将搜索方向和学习率结合在一起考虑的改进算法 (Kingma and Ba, 2015[10]). 它的思路是: 在搜索方向上, 结合动量梯度下降法使用梯度的指数加权; 在学习率上, 使用自适应学习率的调整. 受这一改进梯度下降思想的启发, 我们提出如下 Adam 算法来解决目标函数 $\ell(\eta)$ 极大化问题, 其算法的具体流程如Algorithm 1所示.

Algorithm 1 Adam 算法

Require: $\alpha > 0,\ \epsilon > 0,\ \delta > \epsilon,\ m_0 = 0,\ v_0 = 0,\ t = 0$
Require: $\beta_1,\ \beta_2 \in [0, 1)$
Require: $\eta_1,\ \ell(\eta)$
  while $\delta > \epsilon$ **do**
    $t = t + 1$
    $g_t = -\nabla\ell(\eta_t)$
    $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
    $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
    $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$
    $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$
    $\eta_{t+1} = \eta_t - \frac{\alpha \cdot \hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$
    $\delta = ||\eta_{t+1} - \eta_t||_\infty$
  end while
  return $\hat{\eta} = \eta_t$

在 Adam 算法中, β_1 和 β_2 为指数衰减率, 且要求满足 $\beta_1^2/\sqrt{\beta_2} < 1$. 一般可选取 $\beta_1 = 0.9$, $\beta_2 = 0.999$. 需要指出的是, 所提出的算法中 $\hat{m}_t$ 和 $\hat{v}_t$ 是对 m_t 和 v_t 两项进行了修正. 通过对 m_t 和 v_t 的迭代式累加, 容易得到:

$$m_t = \sum_{k=1}^{t}(1-\beta_1)\beta_1^{t-k}g_k, \quad v_t = \sum_{k=1}^{t}(1-\beta_2)\beta_2^{t-k}g_k^2.$$

由于

$$E(m_t) = (1-\beta_1)\sum_{k=1}^{t}\beta_1^{t-k}E(g_k) = (1-\beta_1^t)E(g_k),$$

$$E(v_t) = (1-\beta_2)\sum_{k=1}^{t}\beta_1^{t-k}E(g_k^2) = (1-\beta_2^t)E(g_k^2).$$

为了使得 $E(m_t) = E(g_k)$, $E(v_t) = E(g_k^2)$, 我们做出以下无偏修正:

$$\begin{aligned}\hat{m}_t &= \frac{m_t}{1-\beta_1^t} = \frac{\sum_{k=1}^{t}(1-\beta_1)\beta_1^{t-k}g_k}{1-\beta_1^t},\\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} = \frac{\sum_{k=1}^{t}(1-\beta_2)\beta_2^{t-k}g_k^2}{1-\beta_2^t}.\end{aligned} \tag{3.3}$$

下面, 我们分析所提出的 Adam 算法的收敛性. 假设以下条件成立.

(C1) η 的参数空间 Θ 是紧集, 且 η_0 在 Θ 的内部, 协变量 X 的取值空间 χ 为紧集.

(C2) 当 $n_V \to \infty$ 时, $n_0/n_V \to \rho_0(> 0)$, $n_k/n_V \to \rho_k(\geq 0)$, $k = 1, \cdots, K$.

(C3) $\sum_{k=1}^{K}\frac{\rho_k}{\rho_0\pi_k^0}\frac{\partial P_k(X,\beta^0)}{\partial\beta_j}$, $j = 1, \cdots, p$ 在 χ 上线性无关. 即: 若 p 维向量 α 使得对于几乎所有的 $X \in \chi$, 满足 $\sum_{j=1}^{p}\alpha_j\sum_{k=1}^{K}\frac{\rho_k}{\rho_0\pi_k^0}\frac{\partial P_k(X,\beta^0)}{\partial\beta_j} = 0$, 则 $\alpha = 0$.

(C4) 存在常数 D 和 D_∞, 使得对任意的 $\eta_1,\ \eta_2 \in \Theta$, 有 $||\eta_1-\eta_2||_2 \le D,\ ||\eta_1-\eta_2||_\infty \le D_\infty$, 其中 $||\cdot||_2$ 和 $||\cdot||_\infty$ 分别表示向量的欧式范数和无穷范数.

(C5) 存在常数 G 和 G_∞, 使得对任意的 t, 有 $||g_t||_2 \le G,\ ||g_t||_\infty \le G_\infty$.

当目标函数是上凸函数时, 我们选择统计量 $R(T)$ 作为收敛性的判定指标:

$$R(T)=\sum_{t=1}^{T}\left[\ell(\hat{\eta})-\ell(\eta_t)\right]. \tag{3.4}$$

当 $T\to\infty$ 时, 若 $R(T)/T\to 0$, 则认为此算法是收敛的(Kingma&Ba, 2015[10]). 特别地, $R(T)$ 随 T 增长得越慢, 算法收敛速度则越快.

用下标 i 表示向量的第 i 个分量, 例如, $g_{k,i}$ 为 $g_k=-\nabla\ell(\eta_k)$ 的第 i 个分量, $\hat{v}_{k,i}$ 为 $\hat{v}_k$ 的第 i 个分量, $k=1,\ldots,t$. 定义 $g_{1:t,i}=(g_{1,i},\cdots,g_{t,i})'$. 关于所提出的 Adam 算法的收敛性, 我们有如下定理.

定理 3.1 在条件(C1)-(C5)成立的情况下, 所提出的 Adam 算法满足如下不等式关系: 对任意的 $T\ge 1$, 有

$$R(T)\le\frac{D^2\sum_{i=1}^{d}\sqrt{T\hat{v}_{T,i}}}{2\alpha(1-\beta_1)}+\frac{\alpha(1+\beta_1)G_\infty\sum_{i=1}^{d}||g_{1:T,i}||_2}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2}+\sum_{i=1}^{d}\frac{D_\infty^2G_\infty\sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2}, \tag{3.5}$$

其中 $d=p+K$ 为参数 η 的维数, $\gamma=\beta_1^2/\sqrt{\beta_2}$, $0<\lambda<1$.

定理 3.1 的证明过程详见附录. 由于不等式(3.5)成立, 故而可得 $\frac{R(T)}{T}=\mathcal{O}(\frac{1}{\sqrt{T}})$, 即: $\lim_{T\to\infty}\frac{R(T)}{T}=0$. 因此, 可以证明所提出的 Adam 算法是收敛的.

进一步地, 针对高维数据的数值计算, 我们改进上述 Adam 算法来提高计算效率. Nesterov动量梯度方法 (Nesterov, 1983[12])是一种改进的动量梯度算法, 它通过动量纠正机制来抑制算法迭代中的震荡, 从而加速收敛. Dozat(2016)[5]在 Adam 算法的基础上融合了这种 Nesterov 动量的思想, 提出了一种 Nadam 算法. 受此启发, 我们将所提出的 Adam 算法中的迭代更新公式:

$$\eta_{t+1}=\eta_t-\frac{\alpha\cdot\hat{m}_t}{\sqrt{\hat{v}_t}+\varepsilon},$$

改进为

$$\eta_{t+1}=\eta_t-\frac{\alpha}{\sqrt{\hat{v}_t}+\varepsilon}\left\{\beta_1\cdot\hat{m}_t+\frac{1-\beta_1}{1-\beta_1^t}\cdot g_t\right\},$$

从而得到我们的 Nadam 算法, 进一步加快数值算法的收敛速度. 根据 Adam 算法的收敛性, 容易获得 Nadam 算法的收敛性. 所提出的 Nadam 算法具体计算步骤总结如下.

Algorithm 2 Nadam 算法

Require: $\alpha > 0,\ \epsilon > 0,\ \delta > \epsilon,\ m_0 = 0,\ v_0 = 0,\ t = 0$
Require: $\beta_1,\ \beta_2 \in [0, 1)$
Require: $\eta_1,\ \ell(\eta)$
  while $\delta > \epsilon$ **do**
    $t = t + 1$
    $g_t = -\nabla\ell(\eta_t)$
    $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
    $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
    $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$
    $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$
    $\eta_{t+1} = \eta_t - \frac{\alpha}{\sqrt{\hat{v}_t+\varepsilon}} \cdot \{\beta_1 \cdot \hat{m}_t + \frac{1-\beta_1}{1-\beta_1^t} \cdot g_t\}$
    $\delta = ||\eta_{t+1} - \eta_t||_\infty$
  end while
  return $\hat{\eta} = \eta_t$

4 模拟研究

本节进行一系列的模拟研究来展示我们所提出的ODS抽样设计下广义线性回归分析中参数估计的数值算法. 首先, 我们考虑二维数据的情形, 假设如下广义线性模型(模型I):

$$Y \sim \lambda \exp\{-\lambda y\} I(y \geq 0), \tag{4.1}$$

其中

$$\lambda = \beta_1 X_1 + \beta_2 X_2.$$

协变量 X_1 由区间 $(1, 2)$ 上的均匀分布生成, X_2 由成功概率为0.5的伯努利分布生成. 设置参数真值为 $\beta_1 = \beta_2 = 0.5$.

对于ODS抽样设计, 假设研究队列包含 N 个相互独立的研究个体. 我们对每个个体测量其因变量的值, 然后根据因变量观测值的分位数点 a_1 和 a_2 做为分割点将队列划分成上、中、下三层. 先随机地抽取一个样本量为 n_0 的SRS样本, 接着从下层和上层分别抽取样本量为 n_1 和 n_3 的补充样本. 为了探究不同分割点对估计的影响, 我们选取了两组分位数点 $(a_1, a_2) = (0.30, 0.70)$ 和 $(0.15, 0.85)$. 设置总样本数 $N = 2000$, 为了研究SRS样本量与补充样本量对估计的影响, 我们考虑分层样本量 $(n_0, n_1, n_3) = (300, 100, 100)$ 和 $(400, 50, 50)$.

为了评估ODS抽样的成本效益, 我们研究参数的两种估计: 一种是所提出的ODS 抽样下的极大半参数经验似然估计 (MSELE), 另一种是基于与ODS抽样具有相同样本量的简单随机样本的极大似然估计 (NAIVE). 分别采用N-R算法、Adam算法和Nadam算法计算这两种估计. 为避免符号混淆, 记算法中指数衰减率为 β_1^*, β_2^*, 取 $\beta_1^* = 0.9,\ \beta_2^* = 0.999$, 基于梯度范数的信息调整参数 ε 的取值, 随机地生成算法初值. 模拟结果包括了估计值相较于参数真值的偏差 (Bias), 估计值的样本标准差 (SD), 标准差估计值的均值 (SE), 以及95%的正态区间覆盖率 (CP). 所有结果均基于1000次独立的模拟结果获得, 如表4.1所示.

表 4.1: 模型 I 中参数 β_1 和 β_2 的模拟结果

				NAIVE				MSELE			
(n_0, n_1, n_3)	(a_1, a_2)	Algorithm		Bias	SD	SE	CP	Bias	SD	SE	CP
(300,100,100)	(0.30,0.70)	NR	$\hat{\beta}_1$	0.0027	0.0318	0.0314	0.942	0.0029	0.0327	0.0317	0.948
			$\hat{\beta}_2$	0.0017	0.0902	0.0896	0.952	0.0018	0.0845	0.0830	0.946
		Adam	$\hat{\beta}_1$	0.0014	0.0317	0.0315	0.940	0.0028	0.0327	0.0317	0.947
			$\hat{\beta}_2$	0.0003	0.0861	0.0893	0.964	0.0018	0.0845	0.0830	0.949
		Nadam	$\hat{\beta}_1$	0.0022	0.0312	0.0315	0.955	0.0028	0.0326	0.0317	0.949
			$\hat{\beta}_2$	0.0014	0.0882	0.0895	0.950	0.0022	0.0827	0.0830	0.952
	(0.15,0.85)	NR	$\hat{\beta}_1$	0.0029	0.0311	0.0314	0.945	0.0021	0.0307	0.0305	0.947
			$\hat{\beta}_2$	0.0051	0.0904	0.0899	0.948	0.0081	0.0755	0.0745	0.942
		Adam	$\hat{\beta}_1$	0.0012	0.0320	0.0313	0.940	0.0019	0.0308	0.0305	0.947
			$\hat{\beta}_2$	0.0057	0.0934	0.0898	0.938	0.0081	0.0755	0.0745	0.942
		Nadam	$\hat{\beta}_1$	0.0004	0.0309	0.0313	0.950	0.0015	0.0309	0.0305	0.944
			$\hat{\beta}_2$	0.0020	0.0890	0.0894	0.946	0.0084	0.0753	0.0744	0.943
(400,50,50)	(0.30,0.70)	NR	$\hat{\beta}_1$	0.0014	0.0317	0.0313	0.945	0.0038	0.0320	0.0314	0.947
			$\hat{\beta}_2$	0.0024	0.0909	0.0895	0.950	-0.0002	0.0870	0.0857	0.938
		Adam	$\hat{\beta}_1$	0.0024	0.0313	0.0313	0.949	0.0038	0.0320	0.0314	0.948
			$\hat{\beta}_2$	0.0006	0.0876	0.0892	0.963	-0.0002	0.0869	0.0857	0.940
		Nadam	$\hat{\beta}_1$	0.0011	0.0316	0.0314	0.951	0.0038	0.0319	0.0314	0.946
			$\hat{\beta}_2$	0.0040	0.0895	0.0894	0.946	-0.0001	0.0861	0.0857	0.942
	(0.15,0.85)	NR	$\hat{\beta}_1$	-0.0001	0.0319	0.0312	0.944	0.0033	0.0296	0.0305	0.957
			$\hat{\beta}_2$	0.0050	0.0900	0.0893	0.950	0.0065	0.0794	0.0800	0.945
		Adam	$\hat{\beta}_1$	0.0000	0.0311	0.0313	0.941	0.0032	0.0297	0.0305	0.957
			$\hat{\beta}_2$	0.0046	0.0922	0.0896	0.930	0.0065	0.0794	0.0800	0.945
		Nadam	$\hat{\beta}_1$	0.0021	0.0313	0.0313	0.951	0.0030	0.0296	0.0305	0.957
			$\hat{\beta}_2$	0.0036	0.0908	0.0898	0.946	0.0067	0.0773	0.0800	0.955

在以上所有考虑的情况下, 关于参数 β_1 和 β_2 的估计都是无偏的, 并且标准误差估计值的均值 (SEs) 很好地估计了估计值的样本标准差 (SDs), 置信区间的覆盖率达到了 95% 的水平. 一方面, ODS 抽样下的估计 MSELE 比简单随机抽样下的估计 NAIVE 更有效. 分位数点取 $(0.15, 0.85)$ 相比于取 $(0.30, 0.70)$ 的估计效率更高, 这说明分割点更接近两端效率更高. 分层样本量取 $(300, 100, 100)$ 相比于取 $(400, 50, 50)$ 的估计效率更高, 这说明在 ODS 样本量固定时, 补充样本被分配的样本量更多效率更高. 另一方面, Adam 算法和 Nadam 算法计算得到的估计结果与 N-R 算法基本一致. 这说明在低维数据情形下, Adam 算法和 Nadam 算法均可以很好地替代 N-R 算法.

然后, 我们考虑高维数据情形, 假设广义线性模型有如下形式(模型II):

$$Y \sim \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\}, \tag{4.2}$$

其中,

$$\mu = \beta_0 + \beta' X,$$

这里, $\beta = (\beta_1, \cdots, \beta_p)$ 是 p 维回归系数, 设置 β 的真值为 $\beta_0 = (0.25 \cdot \mathbf{1}_q', \mathbf{0}_{p-q}')$, 其中 $\mathbf{1}_m$ 和 $\mathbf{0}_m$ 分别表示由1和0组成的 m 维向量. 这里用 $r = q/p$ 表示参数真值 β_0 中非零分量所占的比例. 协变量 X 从 p 维正态分布中生成, 其均值为 $\mathbf{1}_p$, 协方差矩阵 $\Sigma = (\sigma_{ij})_{p\times p}$, $\sigma_{ij} = \rho^{|i-j|}$, $i, j = 1, \cdots, p$. 这里 ρ 表示协变量之间的相关性大小. 对于ODS抽样, 我们设置 $N = 1500$, $(a_1, a_2) = (0.30, 0.70)$ 以及 $(n_0, n_1, n_3) = (300, 100, 100)$. 用所提出的Adam算法和Nadam算法计算ODS抽样下的MSELE估计, 当 $p = 10$, $r = 0.2$, $\rho = 0.5$ 时, 参数估计的模拟结果如表4.2所示. 结果表明所提出的两种算法对于大维数据的MSELE估计均表现良好. 但是, Adam算法的计算时间约是Nadam算法计算时间的2倍 (我们使用的服务器配置如下: CPU: Intel Xeon E5-2630 v2, 12核; CPU主频: 2.6GHz; 内存大小:64GB). 因此, 在实际应用中, 特别是高维数据的分析时, 推荐使用Nadam算法.

表 4.2: 模型II中参数β的模拟结果($p = 10$, $r = 0.2$, $\rho = 0.5$)

	Adam				Nadam			
	Bias	SD	SE	CP	Bias	SD	SE	CP
$\hat{\beta}_1$	0.0031	0.0507	0.0478	0.925	0.0031	0.0507	0.0478	0.925
$\hat{\beta}_2$	0.0004	0.0555	0.0537	0.939	0.0004	0.0555	0.0537	0.940
$\hat{\beta}_3$	0.0029	0.0525	0.0537	0.960	0.0029	0.0525	0.0533	0.960
$\hat{\beta}_4$	0.0004	0.0521	0.0534	0.952	0.0005	0.0522	0.0533	0.951
$\hat{\beta}_5$	-0.0037	0.0516	0.0535	0.953	-0.0038	0.0517	0.0534	0.952
$\hat{\beta}_6$	0.0022	0.0531	0.0534	0.944	0.0022	0.0531	0.0533	0.943
$\hat{\beta}_7$	0.0005	0.0546	0.0535	0.947	0.0005	0.0546	0.0535	0.945
$\hat{\beta}_8$	-0.0010	0.0545	0.0534	0.952	-0.0009	0.0545	0.0532	0.952
$\hat{\beta}_9$	-0.0002	0.0540	0.0535	0.939	-0.0001	0.0540	0.0534	0.938
$\hat{\beta}_{10}$	0.0020	0.0485	0.0478	0.946	0.0020	0.0485	0.0478	0.945

接下来, 在高维数据情形下, 我们比较Nadam算法和N-R算法的计算稳定性. 上一节中提到, N-R算法可能由于信息矩阵的计算奇异性以及算法不收敛而引发计算错误. 同时Nadam算法也可能由于计算不收敛而引发计算错误. 图1展示了1000次模拟中N-R算法和Nadam算法计算错误的次数, 其中 $\rho = 0.8$, $p = 10$, 20, 30, 40, 50 以及 $r = 0.1$. 从图1中可以看到, N-R算法的计算错误次数随维数 p 的升高而增加. 相对应地, 当维数 p 升高时, 所提出的Nadam算法表现非常稳定.

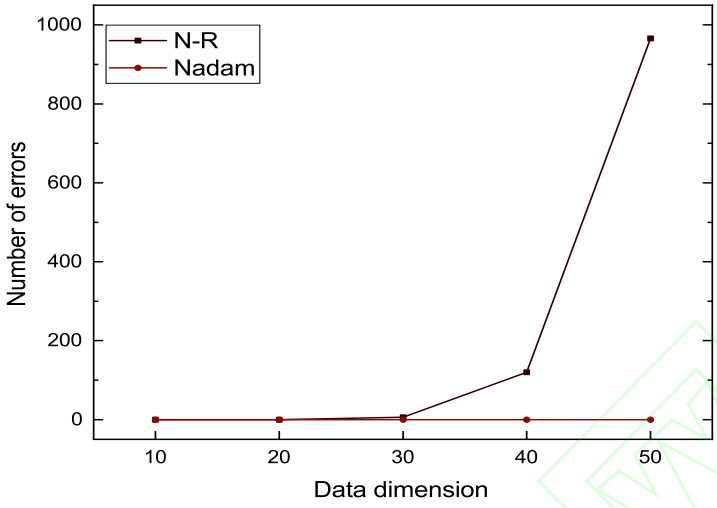

图1: N-R 算法和 Nadam 算法计算出错随 p 变化图

然后, 我们探究协变量间相关性的强度对 Nadam 算法和 N-R 算法的影响. 当 $p = 50$, $r = 0.1$时, ρ 分别取 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8. 模拟 1000 次中 N-R 算法和 Nadam 算法计算出错的次数与 ρ 的关系如图 2 所示. 从图 2 的结果可以看到, 各协变量间的相关性越强, N-R 算法出现计算错误的次数越多, 而本文所提出的 Nadam 算法依然表现优良且非常稳定.

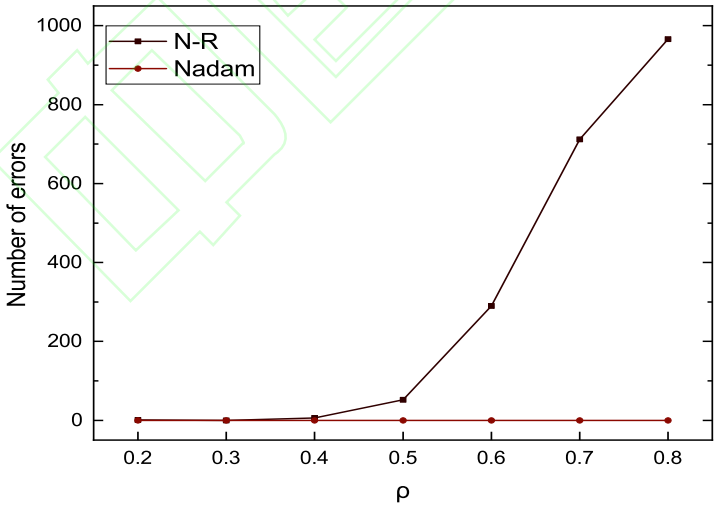

图2: N-R 算法和 Nadam 算法计算出错随 ρ 变化图

最后, 我们探究参数β_0中非零分量的比例r对两种算法稳定性的影响. 当$p = 30$, $\rho = 0.5$时, 分别设置r为0.1, 0.2, 0.3, 0.4, 0.5. 从图3可以看到, 在非稀疏情况下, 所提出的Nadam算法依然表现优良且非常稳定, 而N-R算法出现计算错误的次数随着参数中非零分量的比例r的增加而激增.

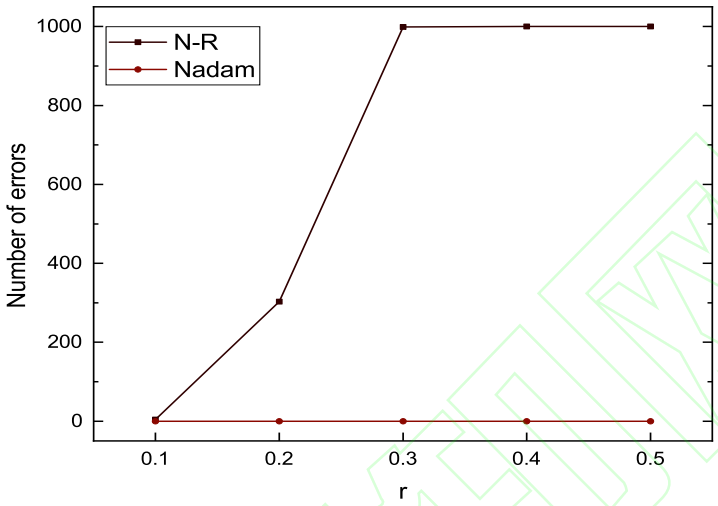

图3: N-R算法和Nadam算法计算出错随r变化图

综上所述, 在数据维数较高, 协变量之间的相关性较强, 或者稀疏性假设不满足的情况下, N-R算法往往会失效, 而我们所提出的Nadam算法表现良好且稳定, 这说明了所提出的Nadam算法在处理高维ODS数据上的优势.

5 实际应用

风能是一种被广泛使用的清洁能源, 风力发电是其主要利用形式之一. 但由于风能的不稳定性, 利用风能发电时需要对电功率进行预测, 以保证电网的稳定性和可靠性, 从而维持电网供电的质量. 利用统计方法根据历史数据建立天气、温度、风向等与输出功率之间的关系, 用实时的测风数据预测实时发电功率是常用的风电预测技术之一.

本节应用所提出的ODS抽样下的估计方法和数值算法来分析一个风电场的实际数据. 该数据来自Chinese Software Developer Network (CSDN) 文库. 数据集包含了新疆某风电场2019年1月和2月每隔15分钟记录的共3648条测风数据. 我们以实际发电功率(MW)为因变量, 考虑如下11个变量: 测风塔10米、30米、50米、70米处的风速(m/s) 以及风向(°)、温度(℃)、气压(hPa)以及湿度(%). 风向数据中以0° 表示正北方向, 90° 表示正东方向. 我们建立了线性回归模型来拟合该数据. 为了评估所提出的ODS抽样机制, 我们人为地从3648条测风数据中随机抽取了1400条作为SRS样本, 基于此SRS样本计算的回归估计做为算法的

初值. 然后分别以实际发电功率的0.30和0.70分位点为分割点, 将所有数据分为三层. 接着, 从实际发电功率较大和较小的两层中各自抽取300条数据作为补充样本. 我们应用ODS抽样下的MSELE推断方法来估计模型回归参数, 使用所提出的Nadam算法实现估计的数值计算. 所得的数据分析结果总结在表5.3中.

表 5.3: 基于ODS抽样的测风数据分析结果

	Est.	SE	p-value
(Intercept)	−233.16	0.0000	< 0.0001
10 m 风速	1.589	0.0043	< 0.0001
30 m 风速	−5.589	0.0048	< 0.0001
50 m 风速	4.194	0.0049	< 0.0001
70 m 风速	13.15	0.0053	< 0.0001
10 m 风向	0.042	0.0028	< 0.0001
30 m 风向	−0.061	0.0024	< 0.0001
50 m 风向	−0.009	0.0023	0.0002
70 m 风向	0.046	0.0025	< 0.0001
温度	−1.898	0.0002	< 0.0001
气压	0.254	0.0040	< 0.0001
湿度	0.228	0.0092	< 0.0001

从表中结果可以看到, 测风塔各位置风速对实时发电功率影响较大, 在测风塔各位置所测风速中，轮毂高度风速即70米风速对发电功率的影响最大, 轮毂高度风速每升高1米每秒, 发电功率升高约13 MW. 其他变量中, 温度和气压的影响较为明显. 温度每升高1 ℃, 实时发电功率降低约1.9 MW. 气压每升高1 hPa, 实时发电功率升高约0.25 MW. 我们计算的同时也发现, 使用ODS抽样设计时, 仅用了全队列中约55%的数据, 得到了与基于全队列数据分析相近的结果. 说明如果采用ODS抽样设计可以起到节省研究成本并提高效率的作用. 结果也表明, 我们提出的Nadam算法表现优良且对初值的选取不敏感.

6 总结

对于大型高维数据的统计研究, 在预算有限时可以使用ODS抽样设计节省成本. ODS抽样设计基于因变量抽取包含信息更多的样本, 减少对某些昂贵协变量观测的开支, 达到节约经费且提高效率的效果. 本文应用广义线性模型来拟合采用ODS抽样设计获取的高维数据. 当模型参数的统计推断方法建立之后, 一个实际中迫切需要解决的问题是估计的数值计算问题. 在数据维度较低时, N-R算法是应用最为广泛的数值算法之一. 但是, 当数据维数较高时, N-R算法很容易遇到一些计算问题. Adam算法是梯度下降算法的改进算法中很为高效的一种. 它将搜索方向和学习率结合在一起考虑, 结合动量梯度下降法使用梯度的指数加

权并使用自适应学习率的调整. Nadam算法在Adam算法的基础上融合了Nesterov动量项来加速Adam算法的收敛. 受这些改进梯度下降思想的启发, 对于高维ODS数据的广义线性回归中目标函数的极大值问题, 我们提出了计算其极大半参数经验似然估计的Adam算法和Nadam算法. 所提出的这些算法避免了高维矩阵及其逆矩阵的计算并抑制了算法在极值处的震荡. 模拟研究和实际数据分析都展示了所提出的算法表现良好、稳定且具有较好的收敛速度. 相较于Adam算法，Nadam算法的计算时间显著降低. 因此, 在实际应用中, 特别是高维数据的分析时, 推荐使用Nadam算法.

在未来工作中, 我们会探索其他一些基于梯度算法思想的改进算法, 例如: Adagrad (Adaptive Gradient) 算法 (Duchi等人, 2011[6]), RMSprop (Root Mean Square Propagation) 算法 (Tieleman &Hinton, 2012[20]), 并比较各种算法的优点和局限性.

附录：引理和定理的证明

引理 6.1 (Yan等人, 2017) 在上述假设条件 $(C1)$-$(C3)$ 下, $\hat{\eta}_V$ 依概率收敛到 η^0, 且

$$\sqrt{n_V}(\hat{\eta}_V - \eta^0) \xrightarrow{d} N(0, \Omega(\eta^0)),$$

其中渐近方差$\Omega(\eta^0) = \Psi^{-1}(\eta^0)\{\Sigma_1(\eta^0)+\Sigma_2(\eta^0)\}\Psi^{-1}(\eta^0)$, $\Psi(\eta)$ 是目标似然函数 (2.5) 的Hessian矩阵的极限矩阵, 且 $\Psi(\eta^0)$ 为正定阵,

$$\Sigma_1(\eta) = \begin{pmatrix} \Sigma_{k=0}^K \rho_k E_k \left\{ -\frac{\partial^2 \ln f_\beta(Y|X)}{\partial\beta\partial\beta'} \right\} & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma_2(\eta) = \Sigma_{k=0}^K \rho_k Var_k \begin{pmatrix} -\frac{\sum_{k=1}^K \{1+\frac{\rho_k}{\rho_0\pi_k}\}\frac{\partial P_k(X,\beta)}{\partial\beta}}{\sum_{k=1}^K \{1+\frac{\rho_k}{\rho_0\pi_k}\}P_k(X,\beta)} \\ \frac{\frac{\rho_1}{\rho_0\pi_1^2}P_1(X,\beta)}{\sum_{k=1}^K \{1+\frac{\rho_k}{\rho_0\pi_k}\}P_k(X,\beta)} - \frac{\rho_1}{\pi_1} \\ \vdots \\ \frac{\frac{\rho_K}{\rho_0\pi_K^2}P_K(X,\beta)}{\sum_{k=1}^K \{1+\frac{\rho_k}{\rho_0\pi_k}\}P_k(X,\beta)} - \frac{\rho_K}{\pi_K} \end{pmatrix}$$

其中 $E_k(h(Y,X)) = \frac{1}{\pi_k^0}\int_\chi \int_{A_k} h(y,x) f_{\beta_0}(y|x)dydQ(x)$.

为证明定理3.1, 基于Kingma&Ba (2015)[10] 的证明思路, 我们先证明一些引理.

引理 6.2

$$\sum_{t=1}^T \sqrt{\frac{g_{t,i}^2}{t}} \le 2G_\infty ||g_{1:T,i}||_2.$$

证明 我们用数学归纳法证明此引理. 当$T = 1$时, 有

$$\sqrt{g_{1,i}^2} \le 2G_\infty ||g_{1,i}||_2.$$

若在$T-1$时不等式成立, 则

$$\begin{aligned}
\sum_{t=1}^{T}\sqrt{\frac{g_{t,i}^2}{t}} &= \sum_{t=1}^{T-1}\sqrt{\frac{g_{t,i}^2}{t}} + \sqrt{\frac{g_{T,i}^2}{T}} \\
&\leq 2G_\infty||g_{1:T-1,i}||_2 + \sqrt{\frac{g_{T,i}^2}{T}} \\
&= 2G_\infty\sqrt{||g_{1:T,i}||_2^2 - g_{T,i}^2} + \sqrt{\frac{g_{T,i}^2}{T}} \\
&\leq 2G_\infty\sqrt{||g_{1:T,i}||_2^2 - g_{T,i}^2 + \frac{g_{T,i}^4}{4||g_{1:T,i}||_2^2}} + \sqrt{\frac{g_{T,i}^2}{T}} \\
&= 2G_\infty(||g_{1:T,i}||_2 - \frac{g_{T,i}^2}{2||g_{1:T,i}||_2}) + \sqrt{\frac{g_{T,i}^2}{T}} \\
&\leq 2G_\infty(||g_{1:T,i}||_2 - \frac{g_{T,i}^2}{2\sqrt{TG_\infty^2}}) + \sqrt{\frac{g_{T,i}^2}{T}} \\
&\leq 2G_\infty||g_{1:T,i}||_2.
\end{aligned}$$

□

引理 6.3

$$\sum_{t=1}^{T}\frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \frac{2G_\infty}{(1-\gamma)^2\sqrt{1-\beta_2}}||g_{1:T,i}||_2.$$

证明 基于式(3.3), 得到:

$$\frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} = \frac{\sqrt{1-\beta_2^t}}{(1-\beta_1^t)^2}\frac{[\sum_{k=1}^{t}(1-\beta_1)\beta_1^{t-k}g_{k,i}]^2}{\sqrt{t\sum_{j=1}^{t}(1-\beta_2)\beta_2^{t-j}g_{j,i}^2}},$$

用数学归纳法易证$(\sum_{k=1}^{T}a_k)^2 \leq T\sum_{k=1}^{T}a_k^2$. 于是, 有

$$\begin{aligned}
\frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} &\leq \frac{t\sqrt{1-\beta_2^t}}{\sqrt{t}(1-\beta_1^t)^2}\frac{\sum_{k=1}^{t}[(1-\beta_1)\beta_1^{t-k}g_{k,i}]^2}{\sqrt{\sum_{j=1}^{t}(1-\beta_2)\beta_2^{t-j}g_{j,i}^2}} \\
&\leq \frac{t\sqrt{1-\beta_2^t}}{\sqrt{t}(1-\beta_1^t)^2}\sum_{k=1}^{t}\frac{[(1-\beta_1)\beta_1^{t-k}g_{k,i}]^2}{\sqrt{(1-\beta_2)\beta_2^{t-k}g_{k,i}^2}} \\
&\leq \frac{t\sqrt{1-\beta_2^t}(1-\beta_1)^2}{\sqrt{t(1-\beta_2)}(1-\beta_1^t)}\sum_{k=1}^{t}\gamma^{t-k}||g_{k,i}||_2 \\
&\leq \frac{t}{\sqrt{t(1-\beta_2)}}\sum_{k=1}^{t}\gamma^{t-k}||g_{k,i}||_2\,.
\end{aligned}$$

接着, 我们逐项迭加, 得到

$$\sum_{t=1}^{T}\frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \sum_{t=1}^{T}\frac{||g_{t,i}||_2}{\sqrt{t(1-\beta_2)}}\sum_{j=0}^{T}t\gamma^j \leq \frac{1}{(1-\gamma)^2\sqrt{1-\beta_2}}\sum_{t=1}^{T}\frac{||g_{t,i}||_2}{\sqrt{t}}\,.$$

由引理6.2, 我们有

$$\sum_{t=1}^{T}\frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}}\leq\frac{2G_\infty}{(1-\gamma)^2\sqrt{1-\beta_2}}||g_{1:T,i}||_2\,.$$

□

定理3.1的证明: 使用Kingma&Ba (2015)[10] 中的证明技巧, 我们取 $\alpha_t=\frac{\alpha}{\sqrt{t}}$, $\beta_{1,t}=\beta_1\lambda^{t-1}$, $\lambda\in(0,1)$. 由于 $\ell(\eta)$ 是上凸函数, 所以满足

$$\ell(\hat{\eta})-\ell(\eta_t)\leq g_t^T(\eta_t-\hat{\eta}),$$

于是统计量 $R(T)$ 满足

$$R(T)=\sum_{t=1}^{T}[\ell(\hat{\eta})-\ell(\eta_t)]\leq\sum_{t=1}^{T}g_t^T(\eta_t-\hat{\eta})=\sum_{t=1}^{T}\sum_{i=1}^{d}g_{t,i}(\eta_{t,i}-\hat{\eta}_i)\,,\tag{A.1}$$

为表示 $g_{t,i}(\eta_{t,i}-\hat{\eta}_i)$, 我们做如下处理:

$$(\eta_{t+1,i}-\hat{\eta}_i)^2=(\eta_{t,i}-\hat{\eta}_i)^2+2(\eta_{t,i}-\hat{\eta}_i)(\eta_{t+1,i}-\eta_{t,i})+(\eta_{t+1,i}-\eta_{t,i})^2,\tag{A.2}$$

由于

$$\eta_{t+1}-\eta_t=-\frac{\alpha_t\hat{m}_t}{\sqrt{\hat{v}_t}}=-\frac{\alpha_t}{1-\beta_1^t}\left\{\frac{\beta_{1,t}m_{t-1}}{\sqrt{\hat{v}_t}}+\frac{(1-\beta_{1,t})g_t}{\sqrt{\hat{v}_t}}\right\},\tag{A.3}$$

将式(A.3)代入式(A.2), 则有:

$$(\eta_{t+1,i}-\hat{\eta}_i)^2=(\eta_{t,i}-\hat{\eta}_i)^2-\frac{2\alpha_t}{1-\beta_1^t}\left\{\frac{\beta_{1,t}m_{t-1,i}}{\sqrt{\hat{v}_{t,i}}}+\frac{(1-\beta_{1,t})g_{t,i}}{\sqrt{\hat{v}_{t,i}}}\right\}(\eta_{t,i}-\hat{\eta}_i)+\left\{\frac{\alpha_t\hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}}}\right\}^2.$$

将 $g_{t,i}(\eta_{t,i}-\hat{\eta}_i)$ 一项分离, 整理得到:

$$\begin{aligned}g_{t,i}(\eta_{t,i}-\hat{\eta}_i)=&\frac{(1-\beta_1^t)\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})}\left\{(\eta_{t,i}-\hat{\eta}_i)^2-(\eta_{t+1,i}-\hat{\eta}_i)^2\right\}+\frac{\alpha_t(1-\beta_1^t)\sqrt{\hat{v}_{t,i}}}{2(1-\beta_{1,t})}\left\{\frac{\hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}}}\right\}^2\\&+\frac{\beta_{1,t}}{1-\beta_{1,t}}\frac{\hat{v}_{t-1,i}^{\frac{1}{4}}}{\sqrt{\alpha_{t-1}}}(\hat{\eta}_i-\eta_{t,i})\sqrt{\alpha_{t-1}}\frac{m_{t-1,i}}{\hat{v}_{t-1,i}^{\frac{1}{4}}},\end{aligned}$$

其中最后一项用 Young 不等式进行放缩, 得到:

$$\begin{aligned}g_{t,i}(\eta_{t,i}-\hat{\eta}_i)\leq&\frac{\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_1)}\left\{(\eta_{t,i}-\hat{\eta}_i)^2-(\eta_{t+1,i}-\hat{\eta}_i)^2\right\}+\frac{\alpha}{2(1-\beta_1)}\frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}}\\&+\frac{\beta_{1,t}(\hat{\eta}_i-\eta_{t,i})^2\sqrt{\hat{v}_{t-1,i}}}{2\alpha_{t-1}(1-\beta_{1,t})}+\frac{\beta_1\alpha}{2(1-\beta_1)}\frac{m_{t-1,i}^2}{\sqrt{(t-1)\hat{v}_{t-1,i}}}.\end{aligned}$$

再将上式代入式(A.1), 并利用引理6.3的结论, 整理即可得到:

$$\begin{aligned}R(T)\leq&\sum_{i=1}^{d}\frac{\sqrt{\hat{v}_{1,i}}}{2\alpha_1(1-\beta_1)}(\eta_{1,i}-\hat{\eta}_i)^2+\sum_{i=1}^{d}\sum_{t=2}^{T}\frac{(\eta_{t,i}-\hat{\eta}_i)^2}{2(1-\beta_1)}\left(\frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t}-\frac{\sqrt{\hat{v}_{t-1,i}}}{\alpha_{t-1}}\right)\\&+\frac{(1+\beta_1)\alpha G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2}\sum_{i=1}^{d}||g_{1:T,i}||_2+\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\beta_{1,t}(\hat{\eta}_i-\eta_{t,i})^2\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})}.\end{aligned}\tag{A.4}$$

根据条件(C4)和(C5), 可整理得

$$\begin{aligned} R(T) \leq & \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T\hat{v}_{T,i}} + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2 \\ & + \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha} \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{\beta_{1,t}}{(1-\beta_{1,t})} \sqrt{t} \,. \end{aligned} \tag{A.5}$$

由于

$$\sum_{t=1}^{T} \frac{\beta_{1,t}}{(1-\beta_{1,t})} \sqrt{t} \leq \sum_{t=1}^{T} \frac{\lambda^{t-1} t}{(1-\beta_1)} \leq \frac{1}{(1-\beta_1)(1-\lambda)^2} \,,$$

代入式(A.5)后, 即可证得目标结果:

$$R(T) \leq \frac{D^2 \sum_{i=1}^{d} \sqrt{T\hat{v}_{T,i}}}{2\alpha(1-\beta_1)} + \frac{\alpha(1+\beta_1)G_\infty \sum_{i=1}^{d} ||g_{1:T,i}||_2}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2} + \sum_{i=1}^{d} \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2}.$$

参 考 文 献

[1] Breslow, N., McNeney, B., Wellner, J. A. (2003). Large sample theory for semi-parametric regression models with two-phase, outcome dependent sampling. *The Annals of Statistics*, **31**: 1110 – 1139.

[2] Boyd, S., Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

[3] Cai, Y., Huang, J., Ning, J., Lee, M. T., Rosner, B., Chen, Y. (2019). Two-sample test for correlated data under outcome-dependent sampling with an application to self-reported weight loss data. *Statistics in Medicine*, **38**: 4999 – 5009.

[4] Ding, J., Lu, T. S., Cai, J., Zhou, H. (2017). Recent progresses in outcome dependent sampling with failure time data. *Lifetime Data Analysis*, **23**: 57 – 82.

[5] Dozat, T. (2016). Incorporating Nesterov momentum into Adam. *ICLR Workshop*, 2016.

[6] Duchi, J., Hazan, E., Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, **12**: 2121 – 2159.

[7] Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**: 1348 – 1360.

[8] Fan, J., Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**: 74 – 99.

[9] Fan, J., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society (Series B)*, **70**: 849 – 911.

[10] Kingma D., Ba J. (2015). Adam: A method for stochastic optimization. *ICLR Workshop*, 2015.

[11] Lange, K. (2004). *Optimization*. Springer.

[12] Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate O(1/k2). *Soviet Mathematics Doklady*, **27**: 372 – 376.

[13] Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural networks : the official journal of the International Neural Network Society*, **12**: 145 – 151.

[14] Qin, G., Zhou, H. (2010). Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome. *Biostatistics*, **12**: 506-520.

[15] Sauer, S., Hedt-Gauthier, B., Rivera-Rodriguez, C., Haneuse, S. (2022). Small sample inference for cluster-based outcome-dependent sampling schemes in resource limited settings: investigating low birthweight in Rwanda. *Biometrics*, **78**: 701 – 715.

[16] Song, R., Zhou, H., Kosorok, M. R. (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, **96**: 221 – 228.

[17] Tan, Z., Qin, G., Zhou, H. (2016). Estimation of a partially linear additive model for data from an outcome-dependent sampling design with a continuous outcome. *Biostatistics*, **17**: 663 – 676.

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

[19] Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, **16**: 385 – 395.

[20] Tieleman, T., Hinton, G.(2012). RMSProp: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural Networks for Machine Learning*, **4**: 26-31.

[21] Yan, S., Ding, J., Liu, Y. (2017). Statistical inference methods and applications of outcome-dependent sampling designs under generalized linear models. *Science China Mathematics*, **60**: 1219 – 1238.

[22] Zhou, H., Weaver, M. A., Qin, J., Longnecker, M. P., Wang, M. C. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 2002, **58**: 413-421.