Vol. 40 (2020)
No. 3

数 学 杂 志
J. of Math. (PRC)

比例风险模型下参数极大似然估计的自适应优化算法及其改进算法

林文强
(武汉大学数学与统计学院, 湖北 武汉 430072)

摘 要: 自适应优化算法可避免很多常用数值算法遭遇的困难, 例如: 高维矩阵求逆问题, 初值选取的问题和算法的收敛问题等等. 因此, 自适应优化算法得到了迅速的发展和广泛的应用, 本文研究了比例风险模型下的自适应优化算法. 首先利用三种自适应优化算法 –Adam 算法、RMSprop 算法、Adagrad 算法求解比例风险模型下的参数估计数值解问题, 获得了自适应算法的计算优良性. 然后, 推广了比例风险模型下的 Adam 算法的研究, 发展了一种改进的 Adam 算法, 进一步提高了算法的计算速度并展现了其计算优势.

关键词: Adam 算法; RMSprop 算法; Adagrad 算法; 比例风险模型
MR(2010) **主题分类号**: 62G05; 62N01 **中图分类号**: O212.2
文献标识码: A **文章编号**: 0255-7797(2020)03-0363-16

1 引言

生存分析主要探究影响事件发生时间的暴露因素, 是统计研究的热点领域之一. 比例风险模型是研究生存数据中应用最为广泛的半参数模型之一. 生存分析涉及领域非常广泛, 例如, 生物学、医学、金融学和管理学等领域. 对比例风险模型中回归参数的估计常常基于其偏似然函数 [1], 通过求解相应的得分方程得到. 由于此得分方程没有解析解, 在实际应用中, 需要通过数值方法来计算待估参数的数值解. Newton-Raphson 算法作为最常用的数值计算方法之一, 有着诸多优点, 例如: 在最优值附近有很快的二次收敛速度; 在求解过程中可以同时给出渐近方差矩阵等. 但是, Newton-Raphson 算法在求解过程中也常常面临一些问题, 例如: 算法涉及二阶导函数矩阵求逆; 可能收敛到局部极值点或者鞍点; 可能不收敛等等 [2]. 因此, 寻找更加稳健的数值计算方法具有重要的理论意义和应用价值.

近年来, 随着深度学习与机器学习领域研究的不断深入, 自适应优化算法得到了迅速发展. 自适应优化算法属于梯度下降优化算法, 其主要特点是: 在迭代过程中引入学习率作为迭代步长, 可沿着负梯度方向不断迭代以接近最优值点, 因而能够在较大程度上节省计算成本和计算空间. 梯度下降优化算法的缺点在于收敛速度较慢, 早期算法如 BGD 算法、SGD 算法等采用固定学习率作为迭代步长, 主要有两方面的问题: 一是学习率的设定及调整问题, 早期算法收敛速度过度依赖学习率的设定. 学习率过小导致收敛速度过慢, 学习率过大导致在到达最优点后过度优化, 进而收敛到次优点或者局部最优点. 二是迭代过程对所有参数采用相同的学习率, 当数据充分稀疏且特征分布不均时, 相同学习率的算法收敛效率和收敛性较差, 在数据分析中被证明是不可取的 (见文献 [3]).

*收稿日期: 2019-05-13 **接收日期:** 2019-09-19
作者简介: 林文强 (1993–), 男, 福建福州, 硕士, 主要研究方向: 统计计算, 生存分析.

针对早期算法的问题, 人们对自适应优化算法的探索主要基于两个方向: 一个方向是基于动量方向的优化算法. 文献 [4] 提出 Momentum 算法, Momentum 算法引入动量的概念, 将当期梯度与前期梯度进行加权求和, 如果两期梯度方向一致, 则加权求和的过程不断累加, 累加过程使得学习率不再依赖于初始学习率的设置. 后期发展的 Nesterov 算法 (Nesterov Momentum) (见文献 [5]), RMSprop 算法 (Root Mean Square Prop) (见文献 [6]) 都是在 Momentum 算法的基础上进行的改进.

另一个方向是基于 L_2 范数下的优化算法. 文献 [3] 提出了 Delta-bar-Delta 方法, 该方法基于一个较为明显的想法: 在给定模型的条件下, 当损失函数对于某参数的偏导持续保持相同的符号, 则增加其更新步长, 反之减少. Delta-bar-Delta 方法是最早自适应优化算法的尝试. 由此文献 [7] 提出了 Adagrad 算法 (Adaptive Gradient Algorithm), Adagrad 算法引进累积梯度平方对学习率进行缩放, 从而为每个参数提供了与自身相关的学习率. 然而, 由于累积梯度平方随着迭代过程递增, Adagrad 算法存在后期学习率衰减的问题. 因此, 文献 [6] 提出了 RMsprop 算法 (Root Mean Square Prop) , RMsprop 算法对累积梯度平方进行了指数滑动平均以丢弃较远的历史数据, 解决了 Adagrad 算法的学习率衰减问题. 进而, 文献 [8] 在 RMSprop 算法的基础上提出了 Adam 算法 (Adaptive Moment Estimation), Adam 算法将指数滑动平均应用到梯度的一阶矩估计当中, 使参数估计更具鲁棒性, 从而实现更快的收敛速度.

据我们所知, 将自适应优化算法应用到比例风险模型的研究尚且较少. 本文将探究比例风险模型下计算参数估计的自适应优化算法及其改进算法, 从分析自适应优化算法在梯度下降良好的收敛性出发, 使用 Adagrad 算法、RMsprop 算法、Adam 算法及其改进算法替代 Newton-Raphson 算法应用于比例风险模型参数估计的求解问题, 探究其合理性和优良性. 在下节中, 先介绍三种自适应优化算法: Adagrad 算法、RMSprop 算法与 Adam 算法.

2 自适应优化算法

本节介绍比例风险模型下, 三种自适应优化算法— Adagrad 算法、RMSprop 算法、Adam 算法求解回归参数的极大似然估计数值解的步骤, 然后, 对自适应优化算法中 Adam 算法进行改进, 使其具有更快的收敛速度.

2.1 比例风险模型下的自适应优化算法

在生存过程中, 用 $\widetilde{T}_i$ 表示第 i 个个体的死亡时间, C_i 表示第 i 个个体的删失时间, $T_i = \mathrm{Min}(\widetilde{T}_i, C_i)$ 表示第 i 个个体的观测时间, $\triangle_i = I(T_i < C_i)$ 表示右删失示性变量, $Z(t)$ 表示 t 时刻的协变量. 文献 [1] 提出如下比例风险率模型

$$\lambda(t|Z) = \lambda_0(t) \exp\left(\beta' Z(t)\right), \tag{2.1}$$

其中 $\lambda_0(t)$ 是基准风险率函数, β 是 p 维待估参数, 其参数空间为 B. 对参数 β 的估计方法, 常见的有矩法估计、最小二乘估计、极大似然似然估计、贝叶斯估计等. 在极大似然估计方法下, 文献 [1] 给出对 β 的统计推断的偏似然函数

$$L(\beta) = \prod_{i=1}^{n} [\frac{e^{\beta' Z_i(T_i)}}{\sum\limits_{l \in R(T_i)} e^{\beta' Z_l(T_i)}}]^{\triangle_i}, \tag{2.2}$$

其中 $R(t)=\{j:T_j\geq t\}$ 表示 t 时刻的风险集, 通常对似然函数取对数

$$l(\beta)=\sum_{i=1}^{n}\triangle_i[\beta' Z_i(T_i)-\log\{\sum_{l\in R(T_i)}e^{\beta' Z_l(T_i)}\}], \tag{2.3}$$

因此 β 的极大似然估计定义为

$$\hat{\beta}=\arg\max_{\beta\in B} l(\beta). \tag{2.4}$$

极大似然估计具有较好的估计性质, 文献 [9] 给出了极大似然估计的相合性证明以及渐近正态性证明. 在实际问题求解过程中, 一般是将极大似然估计的极值问题 (2.4) 转换为得分方程求解问题

$$\nabla l(\beta)=\sum_{i=1}^{n}\triangle_i[Z_i(T_i)-\frac{\sum\limits_{l\in R(T_i)}Z_l(T_i)e^{\beta' Z_l(T_i)}}{\sum\limits_{l\in R(T_i)}e^{\beta' Z_l(T_i)}}]=0. \tag{2.5}$$

由于此得分方程没有解析解, 在实际问题中, 需要采用数值计算方法求解. Newton-Raphon 算法是常用的数值计算方法之一, 在求解过程中需要计算二阶导函数矩阵 $\nabla^2 l(\beta_t)$, 其形式如下

$$\nabla^2 l(\beta_t)=\sum_{i=1}^{n}\triangle_i[\frac{\sum\limits_{l\in R(T_i)}Z(T_i)^{\otimes 2}e^{\beta_t' Z_l(T_i)}}{\sum\limits_{l\in R(T_i)}e^{\beta_t' Z_l(T_i)}}-(\frac{\sum\limits_{l\in R(T_i)}Z_l e^{\beta_t' Z_l(T_i)}}{\sum\limits_{l\in R(T_i)}e^{\beta_t' Z_l(T_i)}})^{\otimes 2}], \tag{2.6}$$

其中 $\otimes$ 表示向量间的运算, 对于向量 x, 定义 $x^{\otimes 2}=xx'$.

因此, 得出比例风险模型下, Newton-Raphon 算法的具体步骤如下

步骤 1 给定初值 β_0 以及计算精度 ϵ;

步骤 2 基于如下迭代公式, 将第 $t-1$ 步估计值 β_{t-1} 更新为第 t 步估计值 β_t

$$\beta_t=\beta_{t-1}-(\nabla^2 l(\beta_t))^{-1}\cdot\nabla l(\beta_t); \tag{2.7}$$

步骤 3 停止准则 $\parallel\beta_t-\beta_{t-1}\parallel<\epsilon$, 即: 不满足时, 令 $t=t+1$, 返回步骤 2; 满足时, 则输出估计值 β_t, 迭代结束.

在实际问题求解中, 二阶导函数矩阵可能是奇异矩阵, 无法求逆进而导致 Newton-Raphon 算法迭代终止. 因而, 需要寻找更稳定的数值计算方法.

首先引入 Adagrad 算法进行求解, Adagrad 算法核心思想是以累积梯度平方 R_t 对迭代步长进行缩小或放大, 即将迭代公式 (2.7) 调整为

$$\beta_t=\beta_{t-1}-\frac{\alpha}{\sqrt{R_t}+\tau}\cdot\nabla l(\beta_t), \tag{2.8}$$

其中 τ 为平滑常数, 其目的是为了防止出现分母为 0 的情况, 一般取为较小的常数. 其中累积梯度平方 R_t 的形式为

$$\begin{aligned} R_1&=\nabla l(\beta_1)^{\odot 2},\\ R_t&=R_{t-1}+\nabla l(\beta_t)^{\odot 2}, \end{aligned} \tag{2.9}$$

其中 $\odot$ 表示元素级别的运算, 对于向量 x, $x^{\odot 2}$ 即对向量里的每个元素进行平方.

由 (2.9) 式不难发现累积梯度平方 R_t 是递增的, 这将造成学习率不断递减. 因而引入 RMsprop 算法. RMSprop 算法利用梯度平方的指数滑动平均过程替换 (2.9) 式的累加过程, 并引入滑动系数 ϕ 控制指数滑动平均的衰减速率, $\phi \in [0,1)$,

$$\begin{aligned} \bar{R}_1 &= (1-\phi)\nabla l(\beta_1)^{\odot 2}, \\ \bar{R}_t &= \phi \bar{R}_{t-1} + (1-\phi)\nabla l(\beta_t)^{\odot 2}. \end{aligned} \tag{2.10}$$

由上可知, $\bar{R}_t$ 是梯度平方 $\nabla l(\beta)^{\odot 2}$ 的指数滑动平均值. RMSprop 算法利用指数滑动平均值 $\bar{R}_t$ 对迭代步长进行调整, 将 (2.7) 式调整为

$$\beta_t = \beta_{t-1} - \frac{\alpha}{\sqrt{\bar{R}_t} + \tau} \cdot \nabla l(\beta_t). \tag{2.11}$$

尽管 RMSprop 算法在求解极值问题时已经达到较快的收敛速度, 但是为了获取更快的收敛速度, 引入 Adam 算法进行求解. 在 Adam 算法下, 定义梯度 $\nabla l(\beta)$ 的有偏一阶矩估计为

$$\begin{aligned} m_1 &= (1-\psi_1)\nabla l(\beta_1), \\ m_t &= \psi_1 m_{t-1} + (1-\psi_1)\nabla l(\beta_t), \end{aligned} \tag{2.12}$$

其中 ψ_1 为控制梯度 $\nabla l(\beta)$ 指数滑动平均过程中衰减速率的超参数, $\psi_1 \in [0,1)$ 默认取值 0.9.

同时定义梯度 $\nabla l(\beta)$ 的有偏二阶矩估计为

$$\begin{aligned} v_1 &= (1-\psi_2)\nabla l(\beta_1)^{\odot 2}, \\ v_t &= \psi_2 v_{t-1} + (1-\psi_2)\nabla l(\beta_t)^{\odot 2}, \end{aligned} \tag{2.13}$$

其中 ψ_2 为控制梯度平方 $\nabla l(\beta)^{\odot 2}$ 指数滑动平均过程中衰减速率的超参数, $\psi_2 \in [0,1)$ 默认取值 0.999.

由 (2.12) 式与 (2.13) 式不难发现, 在 ψ_1, ψ_2 为默认取值的情况下, 在迭代初期, m_t, v_t 是偏向于 0 的, 因此需要进行修正

$$\begin{aligned} \hat{m}_t &= m_t/(1-\psi_1^t), \\ \hat{v}_t &= v_t/(1-\psi_2^t). \end{aligned} \tag{2.14}$$

进而, 将 (2.7) 式替换为

$$\beta_t = \beta_{t-1} - \frac{\alpha}{\sqrt{\hat{v}_t} + \tau} \cdot \hat{m}_t. \tag{2.15}$$

由此, 得出比例风险模型下, Adam 算法的具体步骤如下

步骤 1 给定初值 β_0 以及计算精度 ϵ;

步骤 2 计算梯度的一阶矩估计 $\hat{m}_t$, 其具体形式见 (2.12), (2.14) 式;

步骤 3 计算梯度的二阶矩估计 $\hat{v}_t$, 其具体形式见 (2.13), (2.14) 式;

步骤 4 基于如下迭代公式, 将第 $t-1$ 步估计值 β_{t-1} 更新为第 t 步估计值 β_t,

$$\beta_t = \beta_{t-1} - \frac{\alpha}{\sqrt{\hat{v}_t} + \tau} \cdot \hat{m}_t; \tag{2.16}$$

步骤 5 停止准则 $\| \beta_t - \beta_{t-1} \| < \epsilon$, 即: 不满足时, 令 $t = t+1$, 返回步骤 2; 满足时, 则输出估计值 β_t, 迭代结束.

2.2 自适应优化算法 Adam 的改进

在上一小节中, 介绍了 Adam 算法在比例风险模型下的求解步骤, 在给定初始学习率 α, 平滑常数 τ 时, Adam 算法在第 t 步的迭代步长为

$$\triangle\beta_{\text{Adam}} = \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \tau}. \tag{2.17}$$

结合 (2.13) 式对 $\hat{v}_t$ 进行逐项展开

$$\begin{aligned} \hat{v}_t &= \frac{1}{1-\psi_2^t}\left[\psi_2 v_{t-1} + (1-\psi_2)\nabla l(\beta_t)^{\odot 2}\right] \\ &= \frac{1}{1-\psi_2^t}\left[\psi_2^2 v_{t-2} + \psi_2(1-\psi_2)\nabla l(\beta_{t-1})^{\odot 2} + (1-\psi_2)\nabla l(\beta_t)^{\odot 2}\right] \\ &= \frac{1}{1-\psi_2^t}\left[\psi_2^k v_{t-k} + \psi_2^{k-1}(1-\psi_2)\nabla l(\beta_{t-k+1})^{\odot 2} + \cdots + (1-\psi_2)\nabla l(\beta_t)^{\odot 2}\right], \end{aligned} \tag{2.18}$$

其中 $k = 1, 2, \cdots, t$. 由上式不难发现, 指数滑动平均对近期梯度平方赋予更高的权重, 而远期梯度平方的权重则以指数形式衰减, 从而有效控制了梯度平方的累积过程. 然而, 这种梯度平方的短期记忆能力很可能造成算法收敛到次优的极值点. 发现与文献 [10] 在 ICLR 2018 中正在审核的一篇文章中提出的问题不谋而合.

为了解决梯度平方的短期记忆问题, 引入记号 P_t^A 表示 Adam 算法第 t 步迭代的更新量

$$P_t^A = \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \tau}. \tag{2.19}$$

注意到更新量 P_t^A 与 Adam 算法的迭代步长 (2.17) 式只差一个常数——学习率 α. 结合经典算法 Momentum 算法 (见文献 [4]) 中动量向前的思想, 动量向前的思想是模拟物理里动量的概念, 用积累之前的动量来替代真正的梯度, 记动量符号为 M (Momentum)

$$\begin{aligned} &M_1 = 0, \\ &M_t = u \cdot M_{t-1} + \nabla l(\beta), \end{aligned} \tag{2.20}$$

其中 u 为动量 M 所占的权重.

记 Adam 算法的改进算法为 MAdam 算法 (Modified Adam), 记 MAdam 算法的第 t 步的迭代步长为 P_t^M. 结合动量向前思想 (2.20) 式, 使用 MAdam 算法第 $t-1$ 步的迭代步长与 Adam 算法第 t 步的迭代更新量进行比较, 取二者之间的最大值, 与 Adam 算法第 t 步的迭代更新量进行加权求和, 作为 MAdam 算法第 t 步的迭代步长. 其具体形式为

$$\begin{aligned} &P_1^M = \eta \cdot P_1^A, \\ &P_t^M = \alpha \cdot \text{Max}(P_{t-1}^M, P_t^A) + \eta \cdot P_t^A, \end{aligned} \tag{2.21}$$

其中 α 为学习率, η 为动量因子, 即 Adam 算法更新量 P_t^A 的权重, $\eta \in [0, 1)$.

结合 (2.16) 式, 将 Adam 算法的迭代过程修正为

$$\beta_t = \beta_{t-1} - P_t^M. \tag{2.22}$$

由于 P_t^M 是 P_{t-1}^M 与 P_t^A 最大值与 P_t^A 的加权求和, 因此不难发现在迭代过程中 P_t^M 是以上二者最大值与 P_t^A 的排列组合. 因此分别分析最大值选取的两种情形下的收敛情况, 当 $P_t^A \geq P_{t-1}^M$ 时, MAdam 算法退化为学习率为 $(\alpha+\eta)$ 的 Adam 算法

$$\begin{aligned} P_1^M &= \eta \cdot P_1^A, \\ P_t^M &= (\alpha+\eta) \cdot P_t^A. \end{aligned} \tag{2.23}$$

当 $P_t^A < P_{t-1}^M$ 时, MAdam 算法则演变为衰减速率为 α 的 Adam 算法更新量的指数滑动平均过程

$$\begin{aligned} P_1^M &= \eta \cdot P_1^A, \\ P_t^M &= \alpha \cdot P_{t-1}^M + \eta \cdot P_t^A. \end{aligned} \tag{2.24}$$

由 (2.24) 式, 进行如下推导

$$\begin{aligned} P_t^M &= \alpha \cdot P_{t-1}^M + \eta \cdot P_t^A, \\ P_t^M &= \sum_{j=1}^{t} \alpha^{t-j} \eta \cdot P_j^A. \end{aligned} \tag{2.25}$$

结合 (2.23) 式与 (2.25) 式, 可知 MAdam 算法与 Adam 算法的关系为

$$\begin{aligned} P_1^M &= \eta \cdot P_1^A, \\ P_t^M &= \begin{cases} (\alpha+\eta) \cdot P_t^A, & P_t^A \geq P_{t-1}^M, \\ \sum\limits_{j=1}^{t} \alpha^{t-j} \eta \cdot P_j^A, & P_t^A < P_{t-1}^M. \end{cases} \end{aligned} \tag{2.26}$$

由此, 得到比例风险模型下, MAdam 算法的具体步骤如下

步骤 1 给定初值 β_0 以及计算精度 ϵ;

步骤 2 计算梯度的一阶矩估计 $\hat{m}_t$, 其具体形式见 (2.12), (2.14) 式;

步骤 3 计算梯度的二阶矩估计 $\hat{v}_t$, 其具体形式见 (2.13), (2.14) 式;

步骤 4 计算 Adam 更新量:$P_t^A = \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\tau}$;

步骤 5 计算 MAdam 更新量:$P_t^M = \alpha \cdot \mathrm{Max}(P_t^A, P_{t-1}^M) + \eta \cdot P_t^A$;

步骤 6 基于如下迭代公式, 将第 $t-1$ 步估计值 β_{t-1} 更新为第 t 步估计值 β_t:

$$\beta_t = \beta_{t-1} - P_t^M; \tag{2.27}$$

步骤 7 停止准则 $\| \beta_t - \beta_{t-1} \| < \epsilon$, 即: 不满足时, 令 $t = t+1$, 返回步骤 2; 满足时, 则输出估计值 β_t, 迭代结束.

2.3 自适应优化算法的应用

本节中, 通过数据模拟展示四种自适应优化算法替代 Newton-Raphon 算法的可行性, 展示在不同数据样本量以及不同删失率下, 各优化算法的数值计算时间.

2.4 数值模拟

在给定协变量 Z_1, Z_2 的条件下, 考虑失效时间 $\tilde{T}$ 的风险率函数服从如下形式的比例风险模型

$$\lambda(t|Z_1, Z_2) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2).$$

为了比较各种参数设定下, 自适应优化算法与 Newton-Raphon 算法的参数估计情况, 设定参数真值为 $\beta_1 = 0.693$, $\beta_2 = -0.5$. 协变量 Z_1 从成功概率为 0.5 的 Bernouli 分布中随机生成; 协变量 Z_2 从标准正态分布 $N(0,1)$ 中随机生成. 删失时间 C 服从均匀分布 $U(0,c)$, 设定不同 c 值, 使得 1000 次模拟产生相应的删失率分别为 $\rho = 0.2$, 0.5 或 0.7. 基准风险率 $\lambda_0(t)$ 分别设定为 1 和 $2t$, 由此可分别由风险率为 $\exp\{\beta_1 Z_1 + \beta_2 Z_2\}$ 的指数分布数据, 以及形状参数为 2, 尺度参数为 $1/\sqrt{\exp\{\beta_1 Z_1 + \beta_2 Z_2\}}$ 的 Weibull 分布中随机生成失效时间 $\tilde{T}$. 样本量 n 设置为 100, 200 , 250 或 300.

模拟结果中, Newton-Raphon 算法、Adam 算法、MAdam 算法、RMSprop 算法以及 Adagrad 算法参数估计的估计均值 (Mean)、参数估计的样本标准差 (SD), 标准误差的估计均值 (SE) 和 95% 置信区间覆盖率 (CP) 均由 1000 次独立模拟结果获得. 迭代误差 ϵ 设置为 10^{-5}. 算法的数值计算时间 (Time) 为 1000 次模拟中平均每次计算所耗费的系统计算时间 (单位: 秒), 使用的计算系统配置为:CPU 型号为 Intel Xeon E5-2630 v2, 12 核, CPU 主频为 2.6 GHz, 内存大小为 64 GB.

Adam 算法、MAdam 算法、RMSprop 算法以及 Adagrad 算法的初始学习率设置为 $\alpha = 0.01$, 平滑常数设置为 $\tau = 10^{-8}$.RMSprop 算法的滑动系数使用默认值 $\phi = 0.5$. Adam 算法的超参数使用默认值 $\psi_1 = 0.9$, $\psi_1 = 0.999$, MAdam 算法的超参数与 Adam 算法保持一致, 权重 $\eta = 0.05$. 数值模拟计算结果请见表 1 – 表 4, 各算法计算时间, 请见图 1 – 图 6.

从表 1 – 表 4 的数值计算结果可知, 在以上所有考虑的参数设定情况下, Adam 算法、MAdam 算法、RMSprop 算法得到的极大似然估计均是无偏的, 随着样本量增大, 标准误差的估计均值 SE 和参数估计的样本标准差 SD 逐渐接近, 置信区间覆盖率 CP 也稳定在 0.95 左右, 较为合理. 除了 Adagrad 算法计算偏差过大以外, 其他三种算法得到的参数估计结果与 Newton-Raphon 算法估计结果一致, 这说明自适应优化算法替代 Newton-Raphon 算法的可行性. 在实际应用中, 当 Newton-Raphon 算法出现无法求逆的情况下, 以上三种自适应优化算法是替代 Newton-Raphon 算法的可行方案之一.

进一步, 设置样本量 $n = 50$, 100, 150, 200, 250, 300, 设定删失率分别为 $\rho = 0.2$, 0.5, 0.7. 在计算时间上, 自适应优化算法当中, MAdam 算法的计算效率均高于其他三种自适应优化算法, 这说明对 Adam 算法的改进是可行的, 并有效地缩短了计算时间. 以上几种自适应优化算法在收敛速度上与 Newton-Raphon 算法相比较慢. 但是, 自适应优化算法能克服很多高维数据计算时的困难, 具有其计算优势和应用价值.

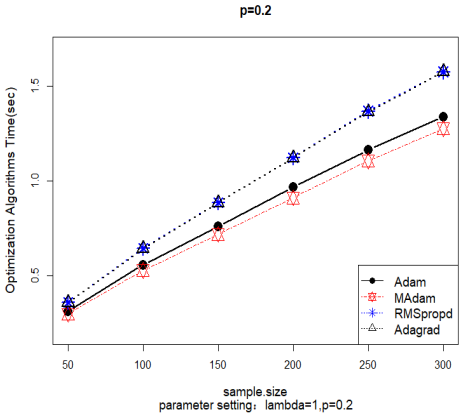

图 1 计算时间比较 ($\lambda_0 = 1, \rho = 0.2$)

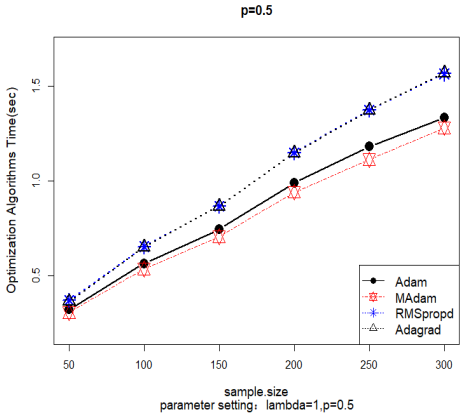

图 2 计算时间比较 ($\lambda_0 = 1, \rho = 0.5$)

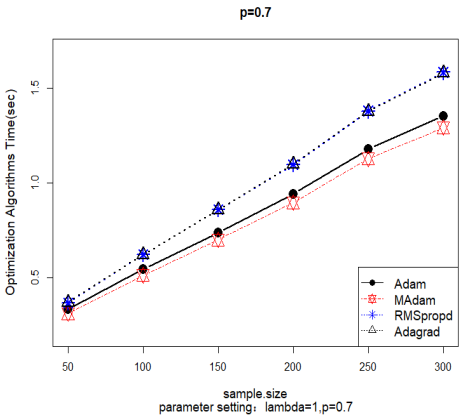

图 3 计算时间比较图 ($\lambda_0 = 1, \rho = 0.7$)

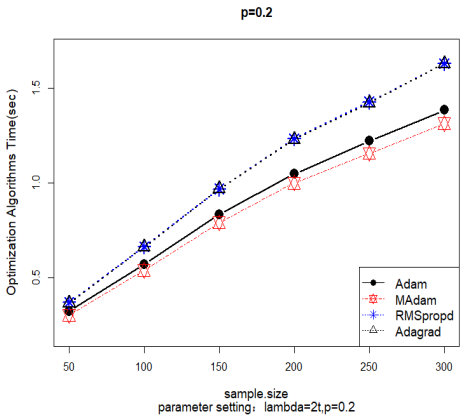

图 4 计算时间比较 ($\lambda_0 = 2t, \rho = 0.2$)

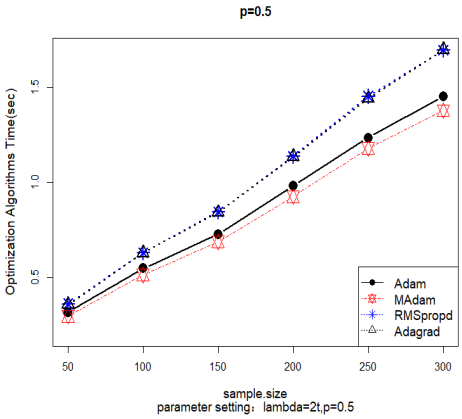

图 5 计算时间比较 ($\lambda_0 = 2t, \rho = 0.5$)

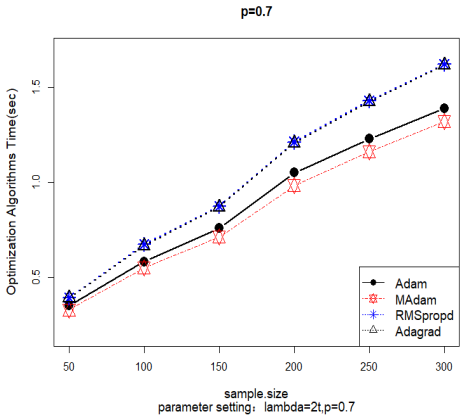

图 6 计算时间比较 ($\lambda_0 = 2t, \rho = 0.7$)

表 1：基于 $\lambda(t|Z_1,Z_2)=\lambda_0(t)\exp(\beta_1Z_1+\beta_2Z_2)$ 的分析结果, 其中 $\lambda_0(t)=1$

				$\beta_1=0.693$				$\beta_2=-0.5$			
n	ρ	Method	Time	Mean	SD	SE	CP	Mean	SD	SE	CP
100	0.2	$\hat{\beta}_{\mathrm{NR}}$	0.014	0.712	0.135	0.137	0.954	−0.514	0.282	0.263	0.944
		$\hat{\beta}_{\mathrm{Adam}}$	0.558	0.712	0.134	0.137	0.953	−0.509	0.271	0.262	0.952
		$\hat{\beta}_{\mathrm{RMSprop}}$	0.646	0.712	0.135	0.137	0.954	−0.514	0.282	0.263	0.944
		$\hat{\beta}_{\mathrm{Adagrad}}$	0.645	0.803	0.053	0.140	0.979	−0.211	0.070	0.240	0.972
		$\hat{\beta}_{\mathrm{MAdam}}$	0.529	0.712	0.135	0.137	0.954	−0.514	0.282	0.263	0.944
	0.5	$\hat{\beta}_{\mathrm{NR}}$	0.014	0.719	0.169	0.166	0.950	−0.513	0.364	0.344	0.942
		$\hat{\beta}_{\mathrm{Adam}}$	0.566	0.718	0.169	0.166	0.950	−0.500	0.335	0.340	0.952
		$\hat{\beta}_{\mathrm{RMSprop}}$	0.655	0.719	0.169	0.166	0.950	−0.513	0.364	0.344	0.942
		$\hat{\beta}_{\mathrm{Adagrad}}$	0.653	0.816	0.078	0.171	0.980	−0.194	0.102	0.304	0.964
		$\hat{\beta}_{\mathrm{MAdam}}$	0.536	0.719	0.169	0.166	0.950	−0.513	0.364	0.344	0.942
	0.7	$\hat{\beta}_{\mathrm{NR}}$	0.014	0.727	0.218	0.208	0.942	−0.543	0.492	0.462	0.949
		$\hat{\beta}_{\mathrm{Adam}}$	0.545	0.725	0.217	0.208	0.942	−0.513	0.424	0.448	0.948
		$\hat{\beta}_{\mathrm{RMSprop}}$	0.623	0.727	0.218	0.208	0.942	−0.543	0.492	0.462	0.949
		$\hat{\beta}_{\mathrm{Adagrad}}$	0.622	0.832	0.105	0.215	0.977	−0.176	0.133	0.392	0.989
		$\hat{\beta}_{\mathrm{MAdam}}$	0.514	0.727	0.219	0.209	0.942	−0.552	0.521	0.461	0.946
200	0.2	$\hat{\beta}_{\mathrm{NR}}$	0.023	0.704	0.091	0.094	0.959	−0.506	0.179	0.183	0.948
		$\hat{\beta}_{\mathrm{Adam}}$	0.967	0.703	0.091	0.094	0.959	−0.505	0.177	0.182	0.955
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.123	0.704	0.091	0.094	0.959	−0.506	0.179	0.183	0.948
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.124	0.793	0.028	0.095	0.982	−0.228	0.024	0.166	0.977
		$\hat{\beta}_{\mathrm{MAdam}}$	0.912	0.704	0.091	0.094	0.959	−0.506	0.179	0.183	0.948
	0.5	$\hat{\beta}_{\mathrm{NR}}$	0.024	0.707	0.116	0.114	0.949	−0.518	0.238	0.239	0.956
		$\hat{\beta}_{\mathrm{Adam}}$	0.990	0.707	0.116	0.114	0.948	−0.514	0.230	0.238	0.959
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.150	0.707	0.116	0.114	0.949	−0.518	0.238	0.239	0.956
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.150	0.800	0.044	0.116	0.978	−0.223	0.042	0.212	0.985
		$\hat{\beta}_{\mathrm{MAdam}}$	0.942	0.707	0.116	0.114	0.949	−0.518	0.238	0.239	0.956
	0.7	$\hat{\beta}_{\mathrm{NR}}$	0.023	0.708	0.144	0.142	0.946	−0.525	0.325	0.316	0.950
		$\hat{\beta}_{\mathrm{Adam}}$	0.944	0.708	0.144	0.142	0.946	−0.514	0.302	0.313	0.969
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.097	0.708	0.144	0.142	0.946	−0.525	0.325	0.316	0.950
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.100	0.808	0.061	0.145	0.974	−0.209	0.075	0.273	0.980
		$\hat{\beta}_{\mathrm{MAdam}}$	0.897	0.708	0.144	0.142	0.947	−0.525	0.325	0.316	0.950

表 2：基于 $\lambda(t|Z_1,Z_2)=\lambda_0(t)\exp(\beta_1Z_1+\beta_2Z_2)$ 的分析结果, 其中 $\lambda_0(t)=1$

				$\beta_1=0.693$				$\beta_2=-0.5$			
n	ρ	Method	Time	Mean	SD	SE	CP	Mean	SD	SE	CP
250	0.2	$\hat{\beta}_{\mathrm{NR}}$	0.028	0.702	0.084	0.083	0.953	−0.501	0.159	0.163	0.955
		$\hat{\beta}_{\mathrm{Adam}}$	1.165	0.702	0.084	0.083	0.953	−0.500	0.158	0.162	0.959
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.370	0.702	0.084	0.083	0.953	−0.501	0.159	0.163	0.955
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.365	0.791	0.026	0.084	0.974	−0.229	0.020	0.148	0.946
		$\hat{\beta}_{\mathrm{MAdam}}$	1.107	0.702	0.084	0.083	0.953	−0.501	0.159	0.163	0.956
	0.5	$\hat{\beta}_{\mathrm{NR}}$	0.028	0.706	0.102	0.101	0.948	−0.510	0.214	0.212	0.947
		$\hat{\beta}_{\mathrm{Adam}}$	1.183	0.706	0.102	0.101	0.949	−0.508	0.210	0.211	0.951
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.376	0.706	0.102	0.101	0.948	−0.510	0.214	0.212	0.947
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.375	0.797	0.036	0.102	0.971	−0.224	0.039	0.189	0.975
		$\hat{\beta}_{\mathrm{MAdam}}$	1.113	0.706	0.102	0.101	0.948	−0.510	0.214	0.212	0.947
	0.7	$\hat{\beta}_{\mathrm{NR}}$	0.029	0.705	0.126	0.126	0.949	−0.530	0.296	0.281	0.950
		$\hat{\beta}_{\mathrm{Adam}}$	1.179	0.705	0.126	0.126	0.950	−0.523	0.280	0.279	0.963
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.381	0.705	0.126	0.126	0.949	−0.530	0.296	0.281	0.950
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.379	0.803	0.052	0.128	0.976	−0.213	0.067	0.243	0.973
		$\hat{\beta}_{\mathrm{MAdam}}$	1.128	0.705	0.126	0.126	0.949	−0.530	0.296	0.281	0.950
300	0.2	$\hat{\beta}_{\mathrm{NR}}$	0.032	0.701	0.075	0.076	0.956	−0.494	0.145	0.148	0.957
		$\hat{\beta}_{\mathrm{Adam}}$	1.339	0.701	0.075	0.076	0.956	−0.494	0.144	0.148	0.959
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.577	0.701	0.075	0.076	0.956	−0.494	0.145	0.148	0.957
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.577	0.790	0.022	0.077	0.974	−0.230	0.019	0.135	0.319
		$\hat{\beta}_{\mathrm{MAdam}}$	1.277	0.701	0.075	0.076	0.956	−0.494	0.145	0.148	0.957
	0.5	$\hat{\beta}_{\mathrm{NR}}$	0.032	0.703	0.092	0.092	0.950	−0.500	0.193	0.193	0.958
		$\hat{\beta}_{\mathrm{Adam}}$	1.335	0.703	0.092	0.092	0.950	−0.499	0.190	0.193	0.961
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.567	0.703	0.092	0.092	0.950	−0.500	0.193	0.193	0.958
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.571	0.794	0.030	0.093	0.979	−0.225	0.035	0.172	0.973
		$\hat{\beta}_{\mathrm{MAdam}}$	1.282	0.703	0.092	0.092	0.950	−0.500	0.193	0.193	0.958
	0.7	$\hat{\beta}_{\mathrm{NR}}$	0.033	0.707	0.116	0.115	0.948	−0.506	0.262	0.255	0.958
		$\hat{\beta}_{\mathrm{Adam}}$	1.355	0.707	0.116	0.115	0.948	−0.502	0.253	0.254	0.964
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.585	0.707	0.116	0.115	0.948	−0.506	0.262	0.255	0.958
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.581	0.801	0.047	0.117	0.969	−0.215	0.058	0.223	0.975
		$\hat{\beta}_{\mathrm{MAdam}}$	1.293	0.707	0.116	0.115	0.948	−0.506	0.262	0.255	0.958

表 3：基于 $\lambda(t|Z_1,Z_2)=\lambda_0(t)\exp(\beta_1 Z_1+\beta_2 Z_2)$ 的分析结果, 其中 $\lambda_0(t)=2t$

				$\beta_1=0.693$				$\beta_2=-0.5$			
n	ρ	Method	Time	Mean	SD	SE	CP	Mean	SD	SE	CP
100	0.2	$\hat{\beta}_{\mathrm{NR}}$	0.014	0.713	0.144	0.139	0.946	−0.509	0.265	0.262	0.948
		$\hat{\beta}_{\mathrm{Adam}}$	0.572	0.712	0.144	0.139	0.946	−0.505	0.258	0.262	0.962
		$\hat{\beta}_{\mathrm{RMSprop}}$	0.665	0.713	0.144	0.139	0.946	−0.509	0.265	0.262	0.948
		$\hat{\beta}_{\mathrm{Adagrad}}$	0.664	0.805	0.058	0.142	0.978	−0.215	0.061	0.242	0.976
		$\hat{\beta}_{\mathrm{MAdam}}$	0.540	0.713	0.144	0.139	0.946	−0.509	0.265	0.262	0.948
	0.5	$\hat{\beta}_{\mathrm{NR}}$	0.013	0.721	0.177	0.169	0.934	−0.521	0.346	0.337	0.952
		$\hat{\beta}_{\mathrm{Adam}}$	0.549	0.720	0.177	0.168	0.935	−0.510	0.323	0.335	0.968
		$\hat{\beta}_{\mathrm{RMSprop}}$	0.631	0.721	0.177	0.169	0.934	−0.521	0.346	0.337	0.952
		$\hat{\beta}_{\mathrm{Adagrad}}$	0.630	0.817	0.081	0.173	0.973	−0.202	0.090	0.303	0.980
		$\hat{\beta}_{\mathrm{MAdam}}$	0.514	0.721	0.177	0.169	0.934	−0.521	0.346	0.337	0.952
	0.7	$\hat{\beta}_{\mathrm{NR}}$	0.015	0.729	0.224	0.213	0.941	−0.537	0.469	0.455	0.963
		$\hat{\beta}_{\mathrm{Adam}}$	0.584	0.727	0.222	0.213	0.941	−0.510	0.416	0.446	0.970
		$\hat{\beta}_{\mathrm{RMSprop}}$	0.675	0.729	0.224	0.213	0.941	−0.537	0.469	0.455	0.963
		$\hat{\beta}_{\mathrm{Adagrad}}$	0.668	0.833	0.107	0.220	0.974	−0.181	0.123	0.395	0.996
		$\hat{\beta}_{\mathrm{MAdam}}$	0.552	0.729	0.224	0.213	0.941	−0.541	0.480	0.455	0.962
200	0.2	$\hat{\beta}_{\mathrm{NR}}$	0.025	0.703	0.097	0.095	0.942	−0.497	0.189	0.182	0.945
		$\hat{\beta}_{\mathrm{Adam}}$	1.049	0.703	0.097	0.095	0.943	−0.496	0.186	0.181	0.947
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.234	0.703	0.097	0.095	0.942	−0.497	0.189	0.182	0.945
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.230	0.793	0.030	0.096	0.980	−0.226	0.032	0.167	0.976
		$\hat{\beta}_{\mathrm{MAdam}}$	1.001	0.703	0.097	0.095	0.943	−0.497	0.189	0.182	0.945
	0.5	$\hat{\beta}_{\mathrm{NR}}$	0.023	0.706	0.116	0.115	0.947	−0.514	0.239	0.233	0.938
		$\hat{\beta}_{\mathrm{Adam}}$	0.984	0.705	0.116	0.115	0.948	−0.511	0.231	0.232	0.942
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.139	0.706	0.116	0.115	0.947	−0.514	0.239	0.233	0.938
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.136	0.798	0.042	0.117	0.977	−0.220	0.048	0.209	0.970
		$\hat{\beta}_{\mathrm{MAdam}}$	0.927	0.706	0.116	0.115	0.947	−0.514	0.239	0.233	0.938
	0.7	$\hat{\beta}_{\mathrm{NR}}$	0.025	0.705	0.147	0.145	0.948	−0.534	0.328	0.314	0.946
		$\hat{\beta}_{\mathrm{Adam}}$	1.052	0.705	0.147	0.145	0.948	−0.524	0.307	0.312	0.956
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.214	0.705	0.147	0.145	0.948	−0.534	0.328	0.314	0.946
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.207	0.807	0.062	0.148	0.973	−0.207	0.081	0.273	0.973
		$\hat{\beta}_{\mathrm{MAdam}}$	0.985	0.705	0.147	0.145	0.948	−0.534	0.328	0.314	0.946

表 4：基于 $\lambda(t|Z_1,Z_2)=\lambda_0(t)\exp(\beta_1Z_1+\beta_2Z_2)$ 的分析结果, 其中 $\lambda_0(t)=2t$

				$\beta_1=0.693$				$\beta_2=-0.5$			
n	ρ	Method	Time	Mean	SD	SE	CP	Mean	SD	SE	CP
250	0.2	$\hat{\beta}_{\mathrm{NR}}$	0.029	0.699	0.085	0.084	0.951	−0.500	0.164	0.162	0.952
		$\hat{\beta}_{\mathrm{Adam}}$	1.223	0.699	0.085	0.084	0.951	−0.499	0.163	0.162	0.953
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.429	0.699	0.085	0.084	0.951	−0.500	0.164	0.162	0.952
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.422	0.791	0.026	0.085	0.978	−0.228	0.026	0.148	0.924
		$\hat{\beta}_{\mathrm{MAdam}}$	1.157	0.699	0.085	0.084	0.951	−0.500	0.164	0.162	0.952
	0.5	$\hat{\beta}_{\mathrm{NR}}$	0.030	0.698	0.105	0.102	0.934	−0.500	0.213	0.207	0.944
		$\hat{\beta}_{\mathrm{Adam}}$	1.236	0.698	0.105	0.102	0.934	−0.499	0.209	0.207	0.950
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.455	0.698	0.105	0.102	0.934	−0.500	0.213	0.207	0.944
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.443	0.795	0.037	0.103	0.972	−0.222	0.046	0.186	0.972
		$\hat{\beta}_{\mathrm{MAdam}}$	1.180	0.698	0.105	0.102	0.934	−0.500	0.213	0.207	0.944
	0.7	$\hat{\beta}_{\mathrm{NR}}$	0.030	0.703	0.133	0.128	0.936	−0.506	0.299	0.278	0.944
		$\hat{\beta}_{\mathrm{Adam}}$	1.230	0.703	0.133	0.128	0.936	−0.499	0.285	0.277	0.950
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.432	0.703	0.133	0.128	0.936	−0.506	0.299	0.278	0.944
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.425	0.804	0.054	0.131	0.969	−0.208	0.078	0.244	0.966
		$\hat{\beta}_{\mathrm{MAdam}}$	1.162	0.703	0.133	0.128	0.936	−0.506	0.299	0.278	0.944
300	0.2	$\hat{\beta}_{\mathrm{NR}}$	0.033	0.698	0.075	0.076	0.954	−0.494	0.155	0.147	0.943
		$\hat{\beta}_{\mathrm{Adam}}$	1.386	0.697	0.075	0.076	0.954	−0.493	0.154	0.147	0.944
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.632	0.698	0.075	0.076	0.954	−0.494	0.155	0.147	0.943
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.631	0.789	0.020	0.077	0.974	−0.229	0.021	0.135	0.347
		$\hat{\beta}_{\mathrm{MAdam}}$	1.315	0.698	0.075	0.076	0.954	−0.494	0.155	0.147	0.943
	0.5	$\hat{\beta}_{\mathrm{NR}}$	0.035	0.700	0.091	0.092	0.955	−0.493	0.193	0.188	0.945
		$\hat{\beta}_{\mathrm{Adam}}$	1.453	0.700	0.091	0.092	0.955	−0.492	0.191	0.188	0.947
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.698	0.700	0.091	0.092	0.955	−0.493	0.193	0.188	0.945
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.699	0.793	0.029	0.094	0.981	−0.224	0.037	0.170	0.966
		$\hat{\beta}_{\mathrm{MAdam}}$	1.379	0.700	0.091	0.092	0.955	−0.493	0.193	0.188	0.945
	0.7	$\hat{\beta}_{\mathrm{NR}}$	0.034	0.701	0.117	0.116	0.952	−0.501	0.256	0.252	0.949
		$\hat{\beta}_{\mathrm{Adam}}$	1.391	0.701	0.117	0.116	0.952	−0.497	0.247	0.251	0.952
		$\hat{\beta}_{\mathrm{RMSprop}}$	1.623	0.701	0.117	0.116	0.952	−0.501	0.256	0.252	0.949
		$\hat{\beta}_{\mathrm{Adagrad}}$	1.620	0.799	0.043	0.118	0.978	−0.216	0.056	0.223	0.973
		$\hat{\beta}_{\mathrm{MAdam}}$	1.322	0.701	0.117	0.116	0.952	−0.501	0.256	0.252	0.949

2.5 实例分析

例 1 (癌症临床实验数据案例) 首先分析一个癌症临床实验的实际案例, 数据来源: CluBear 的公开数据 (熊赳赳教育科技 (北京) 有限公司). 该数据集合收集了 120 个参与某临床试验的癌症病例, 目的是为了发现患者生存时间的相关影响因素, 并对比评估某种新治疗方案的疗效 (同标准治疗方案对比).

该数据集随机抽取某临床试验的 120 个病人, 因变量是病人的生存时间 (单位: 天). 同时数据还提供该病人在试验结束时的生存状态. 那些仍然生存的病人的真实生存时间未知, 因此他们的被记录的生存时间是删失的, 用右删失示性变量 $\delta = 0$ 表示存活, $\delta = 1$ 表示死亡. 解释变量有: 不同治疗方案协变量 $Z_1 = 0$ 表示标准治疗方案, $Z_1 = 1$ 表示新治疗方案; 癌细胞类型协变量 $Z_2 = 1, 2, 3, 4$ 分别表示 A、B、C、D 四种类型的癌细胞; 主治医师临床打分 Z_3 表示医生对病人的综合打分, 连续型变量; 病人的年龄 Z_4(单位: 岁).

表 5：癌症临床实验数据描述性统计分析表

	平均生存时间	$\delta = 1$	$\delta = 0$
Z_1=1(新方案)	139	139.8	128
Z_1=0(旧方案)	117.4	143.4	55.3

建立建立如下比例风险模型评估新治疗方案与生存时间之间的关系

$$\lambda(t|Z_1, Z_2, Z_3, Z_4) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4).$$

模型参数估计的过程使用 Newton-Raphon 算法、Adam 算法以及 MAdam 算法得出参数估计结果, 请见表 6. 可以看出 Adam 算法以及 MAdam 算法与 Newton-Raphon 算法的参数估计结果是一致的. 从数值估计结果来看, 新旧方案没有明显的差别, 协变量 Z_1 的检验 p 值为 0.389. 综上结果表明, 对该癌症案例具有影响的是癌细胞类型与主治医师临床打分, 没有充分的证据可以认定该癌症案例的治疗新方案能够有效的降低患者死亡的风险.

例 2 (天津空气质量数据案例) 本小节, 分析一组空气质量数据, 数据来源: CluBear 的公开数据 (熊赳赳教育科技 (北京) 有限公司). 近年来, 城市雾霾天气在我国频繁出现, 空气质量关乎人民日常正常的生活与劳动, 因此大中型城市空气质量问题已经引起全社会高度关注. 在一份亚洲开发银行与清华大学联合发布的名为《迈向环境可持续的未来中华人民共和国国家环境分析》(见文献 [11]) 报告中指出, 尽管我国政府一直积极运用财政和行政手段治理大气环境污染, 但我国仍是世界上空气污染最严重国家之一. 在我国空气污染最严重的城市排名中, 天津位列前茅.

本文获取了 1809 条数据, 结构如下: 因变量为空气质量指数 (AQI); 协变量 Z_1 表示 PM2.5 指数; Z_2 表示空气质量的 7 个水平: 轻度污染、中度污染、重度污染、严重污染、良、优、无; Z_3 表示 PM10 指数; Z_4 表示空气中 SO_2 含量; Z_5 表示空气中 Co 含量; Z_6 表示空气中 NO_2 含量; Z_7 表示空气中 O3_8h 含量. 建立如下比例风险模型来探究 AQI 影响因素

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \beta_6 Z_6 + \beta_7 Z_7).$$

用 Newton-Raphon 算法、Adam 算法以及 MAdam 算法进行参数估计. 从参数估计的结果来看, 除了协变量 Z_4 (空气中 SO_2 含量) 以外, 其余变量均对 AQI 具有显著影响.

表 6：癌症临床实验数据分析结果

Newton 算法		估计值	标准差	95% 置信区间	p - 值
计算时间 0.136_s	Z_1	0.172	0.200	(-0.22,0.56)	0.389
	Z_2	0.162	0.074	(0.02,0.31)	0.028*
	Z_3	-0.038	0.005	(-0.05,-0.03)	< 0.001*
	Z_4	-0.004	0.009	(-0.02,0.01)	0.651
Adam 算法		估计值	标准差	95% 置信区间	p - 值
计算时间 1.677_s	Z_1	0.174	0.200	(-0.22,0.57)	0.385
	Z_2	0.162	0.074	(0.02,0.31)	0.028*
	Z_3	-0.038	0.005	(-0.05,-0.03)	< 0.001*
	Z_4	-0.004	0.009	(-0.02,0.01)	0.650
MAdam 算法		估计值	标准差	95% 置信区间	p - 值
计算时间 1.346_s	Z_1	0.172	0.200	(-0.22,0.56)	0.388
	Z_2	0.162	0.074	(0.02,0.31)	0.028*
	Z_3	-0.038	0.005	(-0.05,-0.03)	< 0.001*
	Z_4	-0.004	0.009	(-0.02,0.01)	0.651

通过建立比例风险模型, 分析了以上各协变量对空气质量 AQI 的影响, 模型参数的数值计算结果请见表 7. 发现, 三种算法的参数估计结果一致, 与 Adam 算法相比, MAdam 算法的估计结果更接近 Newton-Raphon 算法, 并且计算速度更快. 当待估参数维度增加时, Newton-Raphon 算法涉及高维矩阵的求逆问题. 因此, 相对于 Newton-Raphon 算法, 自适应优化算法在高维参数的估计问题上具有一定的计算优势.

表 7：天津空气质量数据案例分析结果

Newton 算法		估计值	标准差	95% 置信区间	p - 值
（计算时间 11.751_s）	Z_1	-4.842	0.484	$(-5.792, -3.893)$	$< 0.0001^*$
	Z_2	-0.011	0.001	$(-0.013, -0.008)$	$< 0.0001^*$
	Z_3	-0.007	0.001	$(-0.008, -0.005)$	$< 0.0001^*$
	Z_4	0.002	0.001	$(0.000, 0.004)$	0.1055
	Z_5	-0.185	0.071	$(-0.324, -0.046)$	0.0092^*
	Z_6	-0.007	0.002	$(-0.011, -0.003)$	0.0002^*
	Z_7	-0.009	0.001	$(-0.01, -0.007)$	$< 0.0001^*$
Adam 算法		估计值	标准差	95% 置信区间	p - 值
（计算时间 23.229_s）	Z_1	-4.645	0.421	$(-0.529, -0.344)$	$< 0.0001^*$
	Z_2	-0.011	0.001	$(-0.013, -0.008)$	$< 0.0001^*$
	Z_3	-0.007	0.001	$(-0.007, -0.005)$	$< 0.0001^*$
	Z_4	0.002	0.001	$(0.000, 0.005)$	0.1137
	Z_5	-0.182	0.071	$(-0.482, -0.165)$	0.0106^*
	Z_6	-0.007	0.002	$(-0.015, -0.007)$	0.0002^*
	Z_7	-0.009	0.001	$(-0.011, -0.009)$	$< 0.0001^*$
MAdam 算法		估计值	标准差	95% 置信区间	p - 值
（计算时间 20.031_s）	Z_1	-4.804	0.473	$(-2.798, -2.329)$	$< 0.0001^*$
	Z_2	-0.011	0.001	$(-0.013, -0.008)$	$< 0.0001^*$
	Z_3	-0.007	0.001	$(-0.008, -0.005)$	$< 0.0001^*$
	Z_4	0.002	0.001	$(-0.001, 0.004)$	0.0995
	Z_5	-0.184	0.071	$(-0.35, -0.044)$	0.0094^*
	Z_6	-0.007	0.002	$(-0.012, -0.005)$	0.0001^*
	Z_7	-0.009	0.001	$(-0.01, -0.008)$	$< 0.0001^*$

参 考 文 献

[1] David C R. Regression models and life tables (with discussion)[J]. Journal of the Royal Statistical Society, 1972, 34(2): 187–220.

[2] Böhning D, Lindsay B G. Monotonicity of quadratic-approximation algorithms[J]. Annals of the Institute of Statistical Mathematics, 1988, 40(4): 641–663.

[3] Jacobs R A. Increased rates of convergence through learning rate adaptation[J]. Neural networks, 1988, 1(4): 295–307.

[4] Polyak B T. Some methods of speeding up the convergence of iteration methods[J].

USSR Computational Mathematics and Mathematical Physics, 1964, 4(5): 1–17.

[5] Nesterov Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$ $O(1/k2)$[J]. Sov. Math. Doklady, 1983, 27(2): 372–376.

[6] Tieleman T, Hinton G. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude[J]. COURSERA: Neural networks for machine learning, 2012, 4(2): 26–31.

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

[8] Kingma D P, Ba J. Adam: a method for stochastic optimization[J]. https://arxiv.org/abs/1412.6980, 2014.

[9] Andersen P K, Gill R D. Cox's regression model for counting processes: a large sample study[J]. Annal. of Statistics, 1982, 1: 1100–1120.

[10] Dozat T, Manning C D. Deep biaffine attention for neural dependency parsing[J]. https://arxiv.org/abs/1611.01734, 2016.

[11] 张庆丰, 罗伯特, 克鲁. 迈向环境可持续的未来: 中华人民共和国国家环境分析 [M]. 北京: 中国财政经济出版社, 2012.

AUTO-ADAPTED OPTIMIZATION ALGORITHMS AND ITS IMPROVED ALGORITHM FOR PARAMETER MAXIMUM LIKELIHOOD ESTIMATION UNDER THE PROPORTIONAL HAZARDS MODEL

LIN Wen-qiang

(*School of Mathematics and Statistics Wuhan University, Wuhan, 430072, China*)

Abstract: The adaptive optimization algorithm can avoid the difficulties encountered by many commonly used numerical algorithms, such as high-dimensional matrix inversion problem, initial value selection problem and algorithm convergence problem. Therefore, the adaptive optimization algorithm was rapidly developed and widely applied. This paper studies the adaptive optimization algorithm under the proportional risk model. First, three adaptive optimization algorithms, Adam algorithm, RMSprop algorithm and Adagrad algorithm, are used to solve the numerical solution of parameter estimation under the proportional risk model, and the computational superiority of the adaptive algorithm is obtained. Then, the research on the Adam algorithm under the proportional risk model is extended and an improvement is developed. The Adam algorithm further improves the computational speed of the algorithm and demonstrates its computational advantages.

Keywords: Adam algorithms; Rmsprop algorithms; Adagrad algorithms; proportional hazards model

2010 MR Subject Classification: 62G05; 62N01