

基于模拟退火算法的半参数线性回归模型的参数估计

郑国庆, 张国权

(华南农业大学理学院, 广东广州 510642)

摘要: 提出了一种可减少计算量的求解广义半参数极大似然估计的模拟退火算法, 用示例展示新算法在多变量线性回归模型中的应用.

关键词: 广义极大似然估计; 模拟退火算法; 半参数模型; 算法复杂性分析

中图分类号: O213.2 文献标识码: A 文章编号: 1001-411X(2006)02-0115-03

Parameter Estimations of Semi-Parametric Linear Regression

Models Using Simulated Annealing Algorithm

ZHENG Guo-qing, ZHANG Guo-quan

(College of Science, South China Agric. Univ., Guangzhou 510642, China)

Abstract: A simulated annealing algorithm for the generalized semi-parametric maximum likelihood estimator is proposed along with techniques to reduce amount of computation. An example is used to demonstrate the usage of the new algorithm while applying to the multiple linear regression models.

Key words: generalized likelihood; simulated annealing algorithm; semi-parametric models; running cost

考虑在经济学、工程和生命科学中有广泛应用的回归模型

$$Y = \beta'X + \epsilon, \quad (1)$$

其中, β 为 m 维的未知回归系数, ϵ 服从某一未知分布 F_0 . 该模型属于半参数问题, 因为对 (β, ϵ) 再无其他假设. 对回归模型(1)已有多种估计方法, 如最小二乘法、Theil-Sen 估计法^[1-2]、各种 M-估计法^[3-4]、自适应估计法^[5]、L-估计法和 R-估计法^[6]等. 最小二乘法尽管具有优良的统计性质, 却仅当 F_0 为正态分布时才是有效的, 其他估计方法大多依赖于秩统计量, 虽然计算简单, 但都不是有效的, 并且仅当满足一定的正则条件时才是相容的.

最近, Yu 和 Wong^[7-9]研究了半参数模型广义极大似然估计方法. 首先, 定义广义极大似然函数如下:

$$L = \prod_{i=1}^n f(T_i),$$

$$f(t) = F(t) - F(t-) \text{ 且 } F \in \mathcal{F} \quad (2)$$

其中, $T_i = Y_i - \beta'X_i$ 为观测数据, i 代表第 i 次观测. $\mathcal{F} = \{F: F \text{ 为单调增加函数, } F(-\infty)=0 \text{ 且 } F(\infty)=1\}$,

分布函数的广义极大似然估计便是 $\mathcal{F}$ 中使 L 取最大值者. 熟知广义极大似然估计为经验分布函数, 而 (F_0, β) 的半参数极大似然估计(记为 $(\bar{F}_0, \bar{\beta})$)则使

$$L(F, \beta) = \prod_{i=1}^n f(T_i), f(t) = F(t) - F(t-), \quad (3)$$

对所有 $(F, \beta) \in \mathcal{F} \times R^m$ 达到最大, 这里 m 表示 β 的维数.

固定 β , 式(3)中的似然函数在下面的经验分布函数处达到最大

$$\bar{F}_\beta(t) = \frac{1}{n} \sum_{i=1}^n 1\{Y_i - \beta'X_i \leq t\}, \quad (4)$$

其中 $1(\cdot)$ 为示性函数, 任何使 $l(\beta) \equiv L(\bar{F}_\beta, \beta)$ 极大化的 $\bar{\beta}$ 都是 β 的半参数极大似然估计. 在文献[7-8]中, Yu 和 Wong 提出了找出所有半参数极大似然估计的非迭代算法, 并在文献[9]中讨论了渐近性质. 然而, Yu-Wong 算法使用了穷举搜索法, 当回归系数较多时, 计算量急剧增大. 注意到这是搜索空间巨大且存在许多局部极值的非光滑目标函数的全局优化

问题,传统方法往往难以奏效,因此,本文采用模拟退火算法来解决该问题.关于模拟退火算法的介绍及其应用,可参阅文献[10].

1 Yu-Wong 算法的回顾

观察到只要对某些 i, j 存在一个匹配 $T_i(\beta) = T_j(\beta)$, $l(\beta)$ 的值便会增加, Yu 和 Wong 在文献[7-8]中提出了找出所有半参数极大似然估计的非迭代算法:第1步,找出包含所有可能的半参数极大似然估计的集合 B ;第2步,在集合 B 中找出所有使 $l(\beta)$ 极大化的 β ,这些 β 就是要求的半参数极大似然估计.

注解 1.1 上述第1步需求解所有形如 $T_i(\beta) = T_j(\beta)$, 其中 $1 \leq i \neq j \leq n$ 的 m 维线性独立的方程组,以避免遗漏可能的解.下面的结论表明了这项工作计算量非常大.

性质 1.2 假定 $n > 2m$. 则 Yu-Wong 算法的计算量依概率 1 至少为 $C(n, m) = \alpha \times C_n^{2m} \times \frac{(2m)!}{m! 2^m} \times m^3$, 其中 α 为某常量.

证明:讨论计算量下界时,可只考虑最简单的 m 维线性独立方程组情形,即由 $2m$ 个不同的 T_i 构成的形如 $T_i(\beta) = T_j(\beta)$ 的方程组.从 $2m$ 个不同的 T_i 中,有 $\frac{(2m)!}{m! 2^m}$ 种不同的方法将其组成 m 维线性方程组,由于任意一个向量集合依概率 1 线性独立,因此 Yu-Wong 算法需要依概率 1 求解至少 $D(n, m) = C_n^{2m} \times \frac{(2m)!}{m! 2^m}$ 个线性方程组.另外,求解线性方程组的高斯方法的工作量为 $O(m^3)$,极大化 $l(\cdot)$ 的工作量为 $D(n, m)$,因此, Yu-Wong 算法 2.1 的计算量依概率 1 至少为 $D(n, m) \times \alpha \times m^3 = \alpha \times C_n^{2m} \times \frac{(2m)!}{m! 2^m} \times m^3$.

注解 1.3 当 $n \gg m$, $C(n, m) \approx \alpha \times \frac{n^{2m}}{m^{m-3}}$, 故随 m 呈指数增长,即使是对不大的 m ,求解也会变得困难. Yu 和 Wong^[8]曾报道对于 $n = 500, m = 3$ 的一个例子,计算半参数极大似然估计用时超过 1 d.

注解 1.4 Yu 和 Wong^[7-8]还提出了基于估计的模的一些变形和改进算法,然而,这需要对所有估计进行排序,而当 n 和 m 较大时,排序工作量甚至比找到所有可能估计的工作量还要大.由于 Yu-Wong 算法依赖穷举法进行搜索,因此在搜索效率上存在改进的余地,而各种启发式的现代搜索方法正好可以扮演这种角色.

2 求解半参数极大似然估计的模拟退火算法

Yu-Wong 算法不仅工作量大,也没有系统的方法来确定所有的 m 维线性独立方程组,为克服这些困难,本文引入模拟退火算法.其思想是:在搜索过程中,接受 $l(\beta)$ 上升值,并以一定的概率接受下降值避免陷入局部极大值.为此,首先定义状态和邻域概念:

定义 2.1 状态 u 是由观测数据 $T_1, \dots, T_n$ 的下标构成的 $2m$ 元组,记为 $u = (i_1, \dots, i_{2m})$. 状态空间为全体状态的集合.

注解 2.2 状态 u 可确定半参数极大似然估计 β ,因此今后 $l(u)$ 将理解为 $l(\beta)$.

定义 2.3 状态 $u = (i_1, \dots, i_{2m})$ 的邻域是下标与之相差不超过 m 个的所有状态的集合,记为 $\mathcal{N}(u)$.

注解 2.4 邻域关系具有性质:(1)任一状态在它自己的邻域内;(2)若状态 $u \in \mathcal{N}(v)$, 则 $v \in \mathcal{N}(u)$.

注解 2.5 令 $g_{u,v}$ 为由状态 u 转移到 v 的概率,则当 $v \in \mathcal{N}(u)$ 时, $g_{u,v} = 1/n^m$, 否则为零.

注解 2.6 令 V 为状态空间, E 为由邻域诱导的边集,则图 $G(V, E)$ 是强连通的.

最后,定义在温度 t 下,由状态 u 转移到 v 的接受概率 $a_{u,v}(t)$ 为

$$a_{u,v}(t) = \min\{1, e^{\frac{l(v)-l(u)}{t}}\}, \quad (5)$$

显然有:(1)若 $l(u) \leq l(v) \leq l(w)$, 则 $a_{u,w}(t) \leq a_{u,v}(t) \times a_{v,w}(t)$;(2)若 $l(v) \geq l(u)$, 则 $a_{u,v}(t) = 1$, 否则, $0 < a_{u,v}(t) < 1$.

下面给出本文的主要成果.

算法 2.7 找出所有半参数极大似然估计 $\tilde{\beta}$ 的步骤为:第1步,设定初温 t 和初始状态 u ;第2步,从 u 的邻域内随机抽取新状态 u' ;第3步,若 $a_{u,u'}(t)$ 大于或等于 $[0, 1]$ 区间内的所取均匀随机数,则接受 u' 并替换调 u ;第4步,如果抽样稳定准则满足,则转第5步,否则转第2步;第5步,若算法终止条件满足,则输出结果,否则降温并转第2步.

由算法的性质及标准模拟退火算法收敛定理,该算法依概率 1 收敛到全局最优解且不依赖于初始状态的选取.

3 应用与讨论

考虑一个广告收益问题,从 41 份杂志收集的数据画在图 1 中:

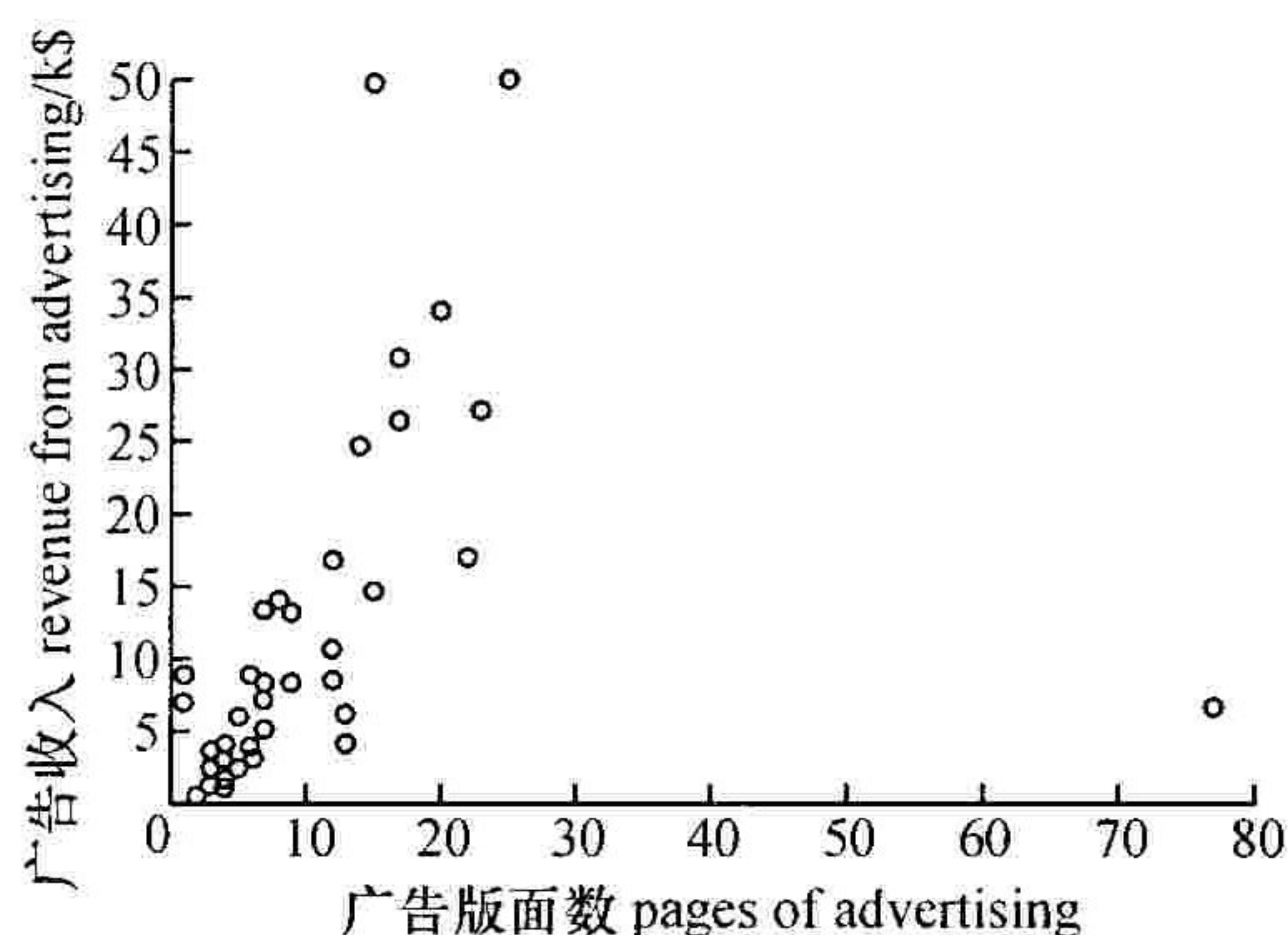


图1 广告收入与广告版面数的关系

Fig. 1 Relationship between revenue v. s. pages of advertising

令 X 为广告版面数, Y 为广告收入,使用如下模型

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \epsilon, \quad (6)$$

最小二乘法(LSE)得到的系数估计为 $\hat{\alpha} = -4.9654$, $\hat{\beta}_1 = 2.1230$, $\hat{\beta}_2 = -0.0255$. Yu-Wong 算法得到的估计为 $\tilde{\alpha} = 2.5868$, $\tilde{\beta}_1 = 0.7630$, $\tilde{\beta}_2 = 0.0174$. 算法2.7运行了100次,得到了与Yu-Wong算法相同的估计,平均迭代次数为478. 考虑到状态空间约有300万种状态,迭代次数相当小. 注意到原始数据中有3个孤立点(15, 49.7), (25, 50.0)和(77, 6.6). 当删除孤立点后,最小二乘法得到的系数变为 $\hat{\alpha} = 0.5692$; $\hat{\beta}_1 = 0.8236$, $\hat{\beta}_2 = 0.0187$ 而Yu-Wong算法和算法2.7则保持不变(表1).

表1 广告收益模型的回归系数估计

Tab. 1 Estimation of regression coefficients of the advertisement model

算法 algorithm	完全数据			删除孤立点数据		
	α	β_1	β_2	α	β_1	β_2
最小二乘法 LSE	-4.9654	2.1230	-0.0255	0.5692	0.8236	0.0187
Yu-Wong 算法 Yu's algorithm	2.5868	0.7630	0.0174	2.5868	0.7630	0.0174
算法 2.7 algorithm 2.7	2.5868	0.7630	0.0174	2.5868	0.7630	0.0174

从表1可以看出,Yu-Wong算法和算法2.7对孤立点数据是稳健的,而最小二乘法则不具备稳健性. 数值试验表明算法2.7比Yu-Wong算法要快得多. 在该例中,尽管数据集相当小($n=41, m=2$),然而,Yu-Wong算法却需要进行高达300万次的方程组求解和极值求解,相反,算法2.7通常在数百次迭代后便收敛到全局最优解.

4 结论

本文提出了一个求解半参数线性回归模型参数估计问题的模拟退火算法,该算法在保留Yu-Wong算法的稳健性等优点的同时,使用了Metropolis接受准则来搜索状态空间里的最优解. 数值试验结果表明新算法的效率比原有算法有明显的提高.

参考文献:

- [1] SEN P K. Estimates of the regression coefficient based on Kendall's tau [J]. JASA, 1968, 63:1379-1389.
- [2] THEIL H. A rank-invariant method of linear and polynomial regression analysis [J], I Proc Kon Ned Akad v Wetensch A, 1950, 53:386-392.
- [3] HUBER P J. Robust estimation of a location parameter [J]. Ann Math Statist, 1964, 35:73-101.

- [4] RITOV Y. Estimation in a linear regression model with censored data [J]. Ann Statist, 1990, 18:303-328.
- [5] BICKEL P J. On adaptive estimation [J]. Ann Statist, 1982, 10:647-671.
- [6] MONTEGOMERY D C, PECK E A. Introduction to linear regression analysis [M]. New York: John Wiley, 1992: 297-316.
- [7] YU Qi-qing, WONG G Y C. Semi-parametric MLE in simple linear regression analysis with interval-censored data [J]. Communication in Statistics-Simulation and Computation, 2003, 32:147-164.
- [8] YU Qi-qing, WONG G Y C. The Semi-parametric MLE in linear regression with right censored data [J]. Journal of Statistical Computation and Simulation, 2003, 73:833-848.
- [9] YU Qi-qing, WONG G Y C. Asymptotic properties of the generalized semi-parametric MLE in linear regression [J]. Statistica Sinica, 2003, 13:311-326.
- [10] ARTS E, KORST J. Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing [M]. New York: Wiley, 1990:250-273.

【责任编辑 李晓卉】