Journal of Mathematical Research with Applications
Mar., 2012, Vol. 32, No. 2, pp. 174–180
DOI:10.3770/j.issn:2095-2651.2012.02.005
Http://jmre.dlut.edu.cn

A Note on the Estimation of Semiparametric Two-Sample Density Ratio Models

Gang YU[1,2,*], **Wei GAO**[3], **Ningzhong SHI**[3]

1. *School of Economics, Huazhong University of Science and Technology, Hubei* 430074*, P. R. China;*
2. *School of Mathematics and Quantitative Economics, Dongbei University of Finance and Economics, Liaoning* 116025*, P. R. China;*
3. *Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Jilin* 130024*, P. R. China*

Abstract In this paper, a semiparametric two-sample density ratio model is considered and the empirical likelihood method is applied to obtain the parameters estimation. A commonly occurring problem in computing is that the empirical likelihood function may be a concave-convex function. Here a simple Lagrange saddle point algorithm is presented for computing the saddle point of the empirical likelihood function when the Lagrange multiplier has no explicit solution. So we can obtain the maximum empirical likelihood estimation (MELE) of parameters. Monte Carlo simulations are presented to illustrate the Lagrange saddle point algorithm.

Keywords empirical likelihood; maximum empirical likelihood estimation (MELE); concave-convex function; Lagrange multiplier; saddle point.

MR(2010) Subject Classification 62F10

1. Introduction

Logistic regression model is commonly used to analyze binary data which arise in studying relationships between diseases and environment or genetic characteristics. Suppose that y is a binary response variable, and x is the covariate vector. The logistic model is $pr(y=1|x)=\exp(\alpha^*+x^{\mathrm{T}}\beta)/\{1+\exp(\alpha^*+x^{\mathrm{T}}\beta)\}\equiv\Lambda(x)$, and the marginal density of x, $f(x)$, is not specified. Define π to be the marginal probability of $y=1$ in the population, i.e., $\pi=\int pr(y=1|x)f(x)\mathrm{d}x$. Let $x_1,\ldots,x_{n_0}$ be from the control group $F(x|y=0), x_{n_0+1},\ldots,x_n$ be from the case group $F(x|y=1)$. Denote the given conditional density functions of x by $f_i(x)=f(x|y=i)=\mathrm{d}F(x|y=i)/\mathrm{d}x, i=0,1$. By Bayes' rule, we have

$$f_1(x)=\Lambda(x)f(x)/\pi,$$

$$f_0(x)=(1-\Lambda(x))f(x)/(1-\pi).$$

Received August 8, 2010; Accepted January 13, 2011
Supported by the National Natural Science Foundation of China (Grant Nos. 10931002; 11071035; 70901016; 71171035) and Excellent Talents Program of Liaoning Educational Committee (Grant No. 2008RC15).

[*] Corresponding author

E-mail address: yugang543@163.com (Gang YU)

We can see that

$$f_1(x)/f_0(x) = \frac{(1-\pi)\Lambda(x)}{\pi(1-\Lambda(x))} = \frac{1-\pi}{\pi}\exp(\alpha^* + x^{\mathrm{T}}\beta) = \exp(\alpha + x^{\mathrm{T}}\beta),$$

where $\alpha = \alpha^* + \log\{(1-\pi)/\pi\}$. So we arrive at a simple semiparametric two-sample density ratio model in which the log ratio of two density functions is linear in data, i.e., after reparameterisation, the assumed logistic regression model is related to a semiparametric two-sample density ratio model. Semiparametric two-sample density ratio models have been studied by Hu et al [1], Guan [2], Zou et al [3], Qin and Zhang [4], Qin [5] and so on, in which the log ratio of two density functions may be nonlinear. Consider a general semiparametric two-sample density ratio model with densities

$$f_0(x), \quad f_1(x) = g(x,\theta)f_0(x), \tag{1}$$

respectively, where $g(x,\theta)$ is a known function, the baseline density f_0 is unknown, and θ is an unknown parameter to be estimated. For model (1), Qin [5] applied maximum empirical likelihood estimation (MELE) to model estimation, and showed the consistency of the maximum empirical likelihood estimator in a neighborhood of the true value θ_0.

In this paper, we consider a special case of (1) in which $g(x,\theta) = \exp\{h^{\mathrm{T}}(x)\theta\}$. Moreover, we discuss the problem of computing maximum empirical likelihood estimation (MELE) in which the empirical loglikelihood function is concave-convex. A concave-convex function is defined as follows:

Definition 1 *Let C and D be subsets of R^m and R^n, respectively, and let $f(x,y)$ be a function from $C \times D$ to $[-\infty,+\infty]$. The function $f(x,y)$ is said to be concave-convex function if $f(x,y)$ is a concave function of $x \in C$ for each $y \in D$ and a convex function of $y \in D$ for each $x \in C$.*

The rest of the paper is organized as follows. In Section 2, for computing issues, we propose a Lagrange saddle point algorithm to obtain the maximum empirical likelihood estimation (MELE). In Section 3, we report some Monte Carlo simulation results. Section 4 concludes the paper.

2. Main results

Suppose we have two independent samples

$$x_1,\dots,x_{n_0} \sim f_0(x), \quad x_{n_0+1},\dots,x_n \sim f_1(x) = \exp\{h^{\mathrm{T}}(x)\theta\}f_0(x), \quad n = n_0+n_1. \tag{2}$$

Under model (2), the loglikelihood is

$$l(\theta,F) = \sum_{i=1}^{n}\log \mathrm{d}F(x_i) + \sum_{i=n_0+1}^{n} h^{\mathrm{T}}(x_i)\theta, \tag{3}$$

where $F = \int f_0$. Here $h^{\mathrm{T}}(\cdot)$ is known vector value function, the link parameter θ and the distribution F are unknown. To maximize $l(\theta,F)$, we discretize F and denote $p_i = \mathrm{d}F(x_i)$ for $i = 1,\dots,n$ as the non-negative jumps with total mass 1, so that

$$l = \sum_{i=1}^{n}\log p_i + \sum_{i=n_0+1}^{n} h^{\mathrm{T}}(x_i)\theta, \tag{4}$$

where

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i\{\exp(h^{\mathrm{T}}(x_i)\theta) - 1\} = 0, \quad p_i \geq 0 \ \ , i = 1, \ldots, n. \tag{5}$$

Similarly to Qin and Lawless [6], for any fixed θ, estimate p_i $(i = 1, \ldots, n)$ by maximization of function (4) subject to constraints (5), which gives

$$p_i = \frac{1}{n}\frac{1}{1 + \rho\{\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1\}}, \tag{6}$$

where ρ is the Lagrange multiplier determined by

$$-\sum_{i=1}^{n} \frac{\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1}{1 + \rho\{\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1\}} = 0. \tag{7}$$

Substituting (6) into (4) gives the empirical loglikelihood function

$$l(\theta, \rho) = -\sum_{i=1}^{n} \log[1 + \rho\{\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1\}] - n\log n + \sum_{i=n_0+1}^{n} h^{\mathrm{T}}(x_i)\theta. \tag{8}$$

Note that $l(\theta, \rho)$ is well defined for $\theta \in \Theta$ and $\rho \in (0, 1)$, where Θ is the parameter space for θ. We can treat θ and ρ as independent parameters in $l(\theta, \rho)$. To obtain a maximum value, taking first-order partial derivative with respect to ρ and θ, we have estimating equations

$$\frac{\partial l}{\partial \rho} = -\sum_{i=1}^{n} \frac{\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1}{1 + \rho\{\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1\}} = 0, \tag{9}$$

$$\frac{\partial l}{\partial \theta} = -\sum_{i=1}^{n} \frac{\rho h(x_i)\exp\{h^{\mathrm{T}}(x_i)\theta\}}{1 + \rho\{\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1\}} + \sum_{i=n_0+1}^{n} h(x_i) = 0, \tag{10}$$

where (9) is equivalent to equation (7). Let $\tilde{\rho}$ and $\tilde{\theta}$ be the solution of the estimating equations (9) and (10). As in Theorem 1 of Qin [5], we can show that $\tilde{\rho}$ converges to n_1/n and $\tilde{\theta}$ is consistent and asymptotically normal in the region $\{\theta| \parallel \theta - \theta_0 \parallel \leq n^{-1/3}\}$, where θ_0 denotes the true value of θ. By implicit differentiation, the second-order partial derivative about ρ and Hessian of $l(\theta, \rho)$ about θ are

$$\frac{\partial^2 l}{\partial \rho^2} = \sum_{i=1}^{n} \frac{[\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1]^2}{[1 + \rho\{\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1\}]^2}, \tag{11}$$

$$\frac{\partial^2 l}{\partial \theta \partial \theta^{\mathrm{T}}} = -\sum_{i=1}^{n} \frac{\rho(\rho - 1)h(x_i)h^{\mathrm{T}}(x_i)\exp\{h^{\mathrm{T}}(x_i)\theta\}}{[1 + \rho\{\exp\{h^{\mathrm{T}}(x_i)\theta\} - 1\}]^2}. \tag{12}$$

Furthermore, it is easy to find that $\frac{\partial^2 l}{\partial \rho^2}$ is positive. Moreover, $\frac{\partial^2 l}{\partial \theta \partial \theta^{\mathrm{T}}}$ is negative definite for fixed ρ in $(0, 1)$, provided that $\sum_{i=1}^{n} h(x_i)h^{\mathrm{T}}(x_i)$ is positive definite. By the Definition 1 in Section 1, the loglikelihood function $l(\theta, \rho)$ is concave-convex. Maximizing $l(\theta, \rho)$ in (θ, ρ) may be unreliable because the function may have a saddle point. A vector pair (θ^*, ρ^*) is said to be a saddle point of $l(\theta, \rho)$ (with respect to maximization in θ and minimization in ρ) if

$$l(\theta, \rho^*) \leq l(\theta^*, \rho^*) \leq l(\theta^*, \rho), \tag{13}$$

for any $\theta \in \Theta$ and $\rho \in (0, 1)$. The following Lemma 1 is extensively discussed in Rockafellar [7, Chap 37]. The essence of the Lemma 1 is that a saddle point of concave-convex function exists under some regularity conditions.

Lemma 1 *Let C and D be non-empty closed bounded convex sets in R^m and R^n, respectively, and let $f(x, y)$ be a continuous finite concave-convex function on $C \times D$. Then $f(x, y)$ has a saddle point with respect to $C \times D$.*

From Lemma 1, we have the following Theorem 1.

Theorem 1 *$l(\theta, \rho)$ has a saddle point (θ^*, ρ^*) which satisfies (13) if $\sum_{i=1}^{n} h(x_i)h^{\mathrm{T}}(x_i)$ is positive definite.*

The above theorem shows that θ^* is just the maximum empirical likelihood estimation (MELE) of θ. Empirical likelihood may pose computational difficulties, see, for example, Owen [8] and Owen [9]. Saddle point algorithm has been studied by some researchers, say, Zhang and Kung [10], He and He [11] and so on. Let $\Phi(\theta, \rho) = \frac{\partial l}{\partial \rho}$. Next we propose a Lagrange saddle point algorithm, based on bisection method as in Bazaraa et al [12], to obtain the saddle point (θ^*, ρ^*).

Lagrange saddle point algorithm:

Initialization Step. Let (a_1, b_1) be the initial interval of ρ, where $a_1 = 0$, $b_1 = 1$. Let m be the allowable final interval of uncertainty. Let n be the smallest positive integer such that

$$(1/2)^n \leq m/(b_1 - a_1).$$

Let $k = 1$ and go to the Main Step.

Main Step.

Step 1. Let $\rho_k = \frac{1}{2}(a_k + b_k)$. Furthermore substituting ρ_k into (10) and solving the score equation (10), we can obtain θ_k by using Newton-Raphson algorithm. Evaluate $\Phi(\theta_k, \rho_k)$. If $\Phi(\theta_k, \rho_k) = 0$, stop; (θ_k, ρ_k) is a saddle point. Otherwise, go to Step 2 if $\Phi(\theta_k, \rho_k) > 0$, and go to Step 3 if $\Phi(\theta_k, \rho_k) < 0$.

Step 2. Let $a_{k+1} = a_k$ and $b_{k+1} = \rho_k$. Go to Step 4.

Step 3. Let $a_{k+1} = \rho_k$ and $b_{k+1} = b_k$. Go to Step 4.

Step 4. If $k = n$, stop, $(\theta_{k+1}, \rho_{k+1})$ is the saddle point; Otherwise, replace k by $k + 1$ and repeat Step 1.

Theorem 2 *The point sequence $\{(\theta_k, \rho_k)\}$ given in the above algorithm converges to a saddle point (θ^*, ρ^*) as $k \to \infty$.*

The proof of Theorem 2 is obvious, so we omit it here.

Remark 1 If the first element of $h^{\mathrm{T}}(x)$ is 1, by differentiating with respect to the first element of θ, we can see easily that the Lagrange multiplier ρ has the explicit solution $\rho = n_1/n$. The point estimator of θ can be easily obtained by solving the following score equation

$$\frac{\partial l}{\partial \theta} = -\sum_{i=1}^{n} \frac{\rho h(x_i)\exp\{h^{\mathrm{T}}(x_i)\theta\}}{1+\rho\{\exp\{h^{\mathrm{T}}(x_i)\theta\}-1\}} + \sum_{i=n_0+1}^{n} h(x_i) = 0, \tag{14}$$

where $\rho = n_1/n$. Qin and Zhang [4], Qin [5] have discussed the above estimating problem.

3. Simulation results

Monte Carlo simulations have been presented to examine the performance of maximum empirical likelihood estimation by using the Lagrange saddle point algorithm. We consider two examples.

Example 1 Take f_0 to be Normal, Exponential, Poisson, where $h^{\mathrm{T}}(x) = (1, x)$. Here let $\theta = (\theta_0, \theta_1)^{\mathrm{T}}$, where θ_0 is known. The Lagrange multiplier ρ has no explicit solution. We report the results for the maximum empirical likelihood estimation of θ_1. Throughout, we show the mean, median and variance of these estimators based on 1000 replications for each design with different n.

$f_0(x)$	$f_1(x)$	n	Mean	Median	Var
$N(0,1)$	$N(2,1)$				
		20	2.188	2.056	0.541
		100	2.032	2.008	0.040
		500	2.010	2.005	0.007
$Ex(1)$	$Ex(2)$				
		20	-1.060	-1.021	0.062
		100	-1.010	-1.000	0.010
		500	-1.000	-0.996	0.002
$Po(3)$	$Po(1)$				
		20	-1.150	-1.115	0.048
		100	-1.106	-1.096	0.007
		500	-1.099	-1.099	0.001

Table 1 Maximum empirical likelihood estimation of θ_1 by using Lagrange sad -dle point algorithm. The true value of θ_1 is 2.0,-1.0,-1.1, respectively.

The point estimator of θ_1 can be also obtained by solving the score equation (14), see the remark 2 of Qin [5].

Example 2 Take f_0 to be Weibull distribution with parameter (2,1), and f_1 to be a Weibull distribution with parameter (4,2), where $h^{\mathrm{T}}(x) = (\log(x), x^2, x^4)$. The true value of $\theta = (\theta_1, \theta_2, \theta_3)^{\mathrm{T}}$ is $(2.0, 1.0, -0.5)^{\mathrm{T}}$. We report the simulation results in Table 2. To save computation, the simulation is repeated 100 times.

Remark 2 In Example 2, Weibull (γ,β) has probability density function given by

$$f(x|\gamma,\beta)=\frac{\gamma}{\beta}x^{\gamma-1}e^{-x^{\gamma}/\beta},\quad 0<x<\infty,\ \gamma>0,\ \beta>0. \tag{15}$$

From Tables 1 and 2, we can see that the maximum empirical likelihood estimation (MELE) by using Lagrange saddle point algorithm is very close to the true value with large sample sizes.

	n	Mean	Median	Var
Estimation of θ_1				
	100	2.249	2.175	0.550
	500	2.033	2.045	0.089
	2000	1.993	1.986	0.023
Estimation of θ_2				
	100	1.057	1.030	0.186
	500	1.019	1.020	0.029
	2000	0.999	1.002	0.006
Estimation of θ_3				
	100	-0.539	-0.512	0.062
	500	-0.516	-0.530	0.009
	2000	-0.499	-0.502	0.002

Table 2 Maximum empirical likelihood estimation of $\theta=(\theta_1,\theta_2,\theta_3)^{\mathrm{T}}$ by using Lagrange saddle point algorithm.

4. Concluding remarks

A semiparametric two-sample density ratio model is proposed. We obtain the maximum empirical likelihood estimation (MELE) by using the Lagrange saddle point algorithm, which can be applicable to more complicated semiparametric two-sample models. From our simulation results, it seems that our proposed Lagrange saddle point algorithm is effective for computing the maximum empirical likelihood estimation (MELE) when the Lagrange multiplier has no explicit solution.

Acknowledgment The authors would like to thank the anonymous referees for helpful suggestions and insightful comments that greatly improved the paper.

References

[1] Zonghui HU, Jing QIN, D. FOLLMANN. *Semiparametric two-sample changepoint model with application to human immunodeficiency virus studies.* J. Roy. Statist. Soc. Ser. C, 2008, **57**(5): 589–607.

[2] Zhong GUAN. *A semiparametric changepoint model.* Biometrika, 2004, **91**(4): 849–862.

[3] Fei ZOU, J. P. FINE, B. S. YANDELL. *On empirical likelihood for a semiparametric mixture model.* Biometrika, 2002, **89**(1): 61–75.

[4] Jing QIN, Biao ZHANG. *A goodness-of-fit test for logistic regression models based on case-control data.* Biometrika, 1997, **84**(3): 609–618.

[5] Jing QIN. *Inferences for case-control and semiparametric two-sample density ratio models*. Biometrika, 1998, **85**(3): 619–630.

[6] Jing QIN, J. LAWLESS. *Empirical likelihood and general estimating equations*. Ann. Statist., 1994, **22**(1): 300–325.

[7] R. T. ROCKAFELLAR. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1972.

[8] A. B. OWEN. *Empirical likelihood ratio confidence intervals for a single functional*. Biometrika, 1988, **75**(2): 237–249.

[9] A. B. OWEN. *Empirical likelihood ratio confidence regions*. Ann. Statist., 1990, **18**(1): 90–120.

[10] Suochun ZHANG, Yu KANG. *A Lagrange-saddle point algorithm for computing periodic solutions*. Math. Numer. Sin., 1996, **18**(2): 199–206.

[11] Bingsheng HE, Xuchu HE. *A class of saddle point algorithm for convex programming*. Math. Numer. Sin., 1991, **13**: 51–57.

[12] M. S. BAZARAA, H. D. SHERALI, C. M. SHETTY. *Nonlinear Programming: Theory and Algorithms*. Wiley-Interscience, 2006.