Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 537474, 15 pages
doi:10.1155/2012/537474

Research Article

Testing Homogeneity in a Semiparametric Two-Sample Problem

Yukun Liu,[1] Pengfei Li,[2] and Yuejiao Fu[3]

[1] *Department of Statistics and Actuarial Science, School of Finance and Statistics, East China Normal University, Shanghai 200241, China*

[2] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1*

[3] *Department of Mathematics and Statistics, York University, Toronto, ON, Canada M3J 1P3*

Correspondence should be addressed to Yuejiao Fu, yuejiao@mathstat.yorku.ca

Received 18 November 2011; Accepted 24 January 2012

Academic Editor: Yongzhao Shao

We study a two-sample homogeneity testing problem, in which one sample comes from a population with density $f(x)$ and the other is from a mixture population with mixture density $(1-\lambda)f(x)+\lambda g(x)$. This problem arises naturally from many statistical applications such as test for partial differential gene expression in microarray study or genetic studies for gene mutation. Under the semiparametric assumption $g(x) = f(x)e^{\alpha+\beta x}$, a penalized empirical likelihood ratio test could be constructed, but its implementation is hindered by the fact that there is neither feasible algorithm for computing the test statistic nor available research results on its theoretical properties. To circumvent these difficulties, we propose an EM test based on the penalized empirical likelihood. We prove that the EM test has a simple chi-square limiting distribution, and we also demonstrate its competitive testing performances by simulations. A real-data example is used to illustrate the proposed methodology.

1. Introduction

Let $x_1, \ldots, x_{n_0}$ be a random sample from a population with distribution function F, and let $y_1, \ldots, y_{n_1}$ be a random sample from a population with distribution function H. Testing whether the two populations have the same distribution, that is, $H_0 : F = H$ versus $H_1 : F \neq H$, with both F and H completely unspecified, will require a nonparametric test. Since $H_1 : F \neq H$ is a very broad hypothesis, many times one may want to consider some more specified alternative, for example, the two populations only differ in location. In the present paper, we will consider a specified alternative in which one of the two samples has a mixture structure. More specifically, we have

$$x_1, \ldots, x_{n_0} \overset{\text{i.i.d.}}{\sim} f(x), \qquad y_1, \ldots, y_{n_1} \overset{\text{i.i.d.}}{\sim} h(y) = (1-\lambda)f(y) + \lambda g(y), \tag{1.1}$$

where $f(x) = dF(x)/dx$, $g(y) = dG(y)/dy$, $h(y) = dH(y)/dy$, and $\lambda \in (0,1)$ is an unknown parameter sometimes called contamination proportion. The problem of interest is to test H_0 : $f = h$ or equivalently $\lambda = 0$. This particular two-sample problem arises naturally in a variety of statistical applications such as test for partial differential gene expression in microarray study, genetic studies for gene mutation, case-control studies with contaminated controls, or the test of a treatment effect in the presence of nonresponders in biological experiments (see Qin and Liang [1] for details).

If no auxiliary information is available, this is merely the usual two-sample goodness-of-fit problem. There has been extensive literature on it; see Zhang [2] and references therein. However, these tests are not suitable for the specific alternative with a mixture structure as they might be inferior comparing with methods that are designed for the specific alternative. In this paper, we will propose an empirical likelihood-based testing procedure for this specific mixture alternative under Anderson's semiparametric assumption [3]. Motivated by the logistic regression model, the semiparametric assumption proposed by Anderson [3] links the two distribution functions F and G through the following equation:

$$\log \frac{g(x)}{f(x)} = \alpha + \beta x, \tag{1.2}$$

where α and β are both unknown parameters. There are many examples where the logarithm of the density ratio is linear in the observations.

Example 1.1. Let F and G be the distribution functions of Binomial (m, p_1) and Binomial (m, p_2), respectively. We refer the densities f and g to the probability mass functions corresponding to F and G, respectively. Then,

$$\log \frac{g(x)}{f(x)} = m \log\left\{\frac{1-p_2}{1-p_1}\right\} + \log\left\{\frac{p_2(1-p_1)}{p_1(1-p_2)}\right\} x. \tag{1.3}$$

Example 1.2. Let F be the distribution function of $N(\mu_1, \sigma^2)$ and G the distribution function of $N(\mu_2, \sigma^2)$. Then,

$$\log \frac{g(x)}{f(x)} = \frac{1}{2\sigma^2}\left(\mu_1^2 - \mu_2^2\right) + \frac{1}{\sigma^2}(\mu_2 - \mu_1)x. \tag{1.4}$$

In practice, one may need to apply some sort of transformation to the data (e.g., logarithm transformation) in order to justify the use of the semiparametric model assumption (1.2).

Example 1.3. Let F and G be the distribution functions of $\log N(\mu_1, \sigma^2)$ and $\log N(\mu_2, \sigma^2)$, respectively. It is clear that the density ratio is a linear function of the log-transformed data:

$$\log \frac{g(x)}{f(x)} = \frac{1}{2\sigma^2}\left(\mu_1^2 - \mu_2^2\right) + \frac{1}{\sigma^2}(\mu_2 - \mu_1)\log x. \tag{1.5}$$

Example 1.4. Let F and G be the distribution functions of Gamma (m_1, θ) and Gamma (m_2, θ), respectively. In this case,

$$\log \frac{g(x)}{f(x)} = \log\left\{ \frac{\Gamma(m_1)}{\Gamma(m_2)} \right\} + (m_1 - m_2) \log \theta + (m_2 - m_1) \log x. \tag{1.6}$$

The semiparametric modeling assumption (1.2) is very flexible and has the advantage of not putting any specific restrictions on the functional form of f. Under this assumption, various approaches have been proposed to test homogeneity in the two-sample problem (see [1, 4, 5] and references therein). This paper adds to this literature by introducing a new type of test statistics which are based on the empirical likelihood [6, 7].

The empirical likelihood (EL) is a nonparametric likelihood method which has many nice properties paralleling to the likelihood methods, for example, it is range-preserving, transform-respect, Bartlett correctable, and a systematic approach to incorporating auxiliary information [8–11]. In general, if the parameters are identifiable, the empirical likelihood ratio (ELR) test has a chi-square limiting distribution under null hypothesis. However, for the aforementioned testing problem, the parameters under H_0 are not identifiable, which results in an intractable null limiting distribution for the ELR test. To circumvent this problem, we would add a penalty to the log EL to penalize λ being too close to zero. Working like a soft threshold, the penalty makes the parameters roughly identifiable. Intuitively, the penalized (or modified) ELR test should restore the usual chi-square limiting distribution. Unfortunately two things hinder the direct use of the penalized ELR test. One is that, to the best of our knowledge, there is no feasible algorithm to compute the penalized ELR test statistic. The other one is that there has been no research on the asymptotic properties of the penalized ELR test. Therefore, one cannot obtain critical values for the penalized ELR test regardless through simulations or an asymptotic reference distribution. We find that the EM test [12, 13] based on the penalized EL is a nice solution to the testing problem.

The remainder of this paper is organized as follows. In Section 2, we introduce the ELR and the penalized ELR. The penalized EL-based EM test is given in Section 3. A key computational issue of the EM test is discussed in Section 4. Sections 5 and 6 contain a simulation study and a real-data application, respectively. For clarity, all proofs are postponed to the appendix.

2. Empirical Likelihood

Let $\{t_1, \ldots, t_{n_0}, t_{n_0+1}, \ldots, t_n\} = \{x_1, \ldots, x_{n_0}, y_1, \ldots, y_{n_1}\}$ denote the combined two-sample data, where $n = n_0 + n_1$. Under Anderson's semiparametric assumption (1.2), the likelihood of two-sample data (1.1) is

$$L = \prod_{i=1}^{n_0} dF(t_i) \prod_{j=n_0+1}^{n} \left[1 - \lambda + \lambda e^{\alpha+\beta t_j}\right] dF(t_j). \tag{2.1}$$

Let $p_h = dF(t_h)$, $h = 1, \ldots, n$. The EL is just the likelihood L with constraints $p_h \geq 0$, $\sum_{h=1}^{n} p_h = 1$ and $\sum_{h=1}^{n} p_h(e^{\alpha+\beta t_h} - 1) = 0$. The corresponding log-EL is

$$l = \sum_{h=1}^{n} \log p_h + \sum_{j=1}^{n_1} \log\left[1 - \lambda + \lambda e^{\alpha+\beta y_j}\right]. \tag{2.2}$$

We are interested in testing

$$H_0 : \lambda = 0 \quad \text{or} \quad (\alpha, \beta) = (0, 0). \tag{2.3}$$

Under the null hypothesis, the constraint $\sum_{h=1}^{n} p_h(e^{\alpha+\beta t_h} - 1) = 0$ will always hold and $\sup_{H_0} l = -n \log n$. Under alternative hypothesis, for any fixed (λ, α, β), maximizing l with respect to p_h's leads to the log-EL function of (λ, α, β):

$$l(\lambda, \alpha, \beta) = -\sum_{h=1}^{n} \log\left[1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)\right] - n \log n + \sum_{j=1}^{n_1} \log\left[1 - \lambda + \lambda e^{\alpha+\beta y_j}\right], \tag{2.4}$$

where ξ is the solution to the following equation:

$$\sum_{h=1}^{n} \frac{e^{\alpha+\beta t_h} - 1}{1 + \xi(e^{\alpha+\beta t_h} - 1)} = 0. \tag{2.5}$$

Hence, the EL ratio function $R(\lambda, \alpha, \beta) = 2\{l(\lambda, \alpha, \beta) + n \log n\}$ and the ELR is denoted as $R = \sup R(\lambda, \alpha, \beta)$.

The null hypothesis H_0 holds for $\lambda = 0$ regardless of (α, β), or $(\alpha, \beta) = (0, 0)$ regardless of λ. This implies that the parameter (λ, α, β) is not identifiable under H_0, resulting in rather complicated asymptotic properties of the ELR. One may consider the modified or penalized likelihood method [14] and define the penalized log-EL function $pl(\lambda, \alpha, \beta) = l(\lambda, \alpha, \beta) + \log(\lambda)$. Accordingly the penalized EL ratio function is

$$\begin{aligned} pR(\lambda, \alpha, \beta) &= 2\{pl(\lambda, \alpha, \beta) - pl(1, 0, 0)\} \\ &= -2\sum_{h=1}^{n} \left[1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)\right] \\ &\quad + 2\sum_{j=1}^{n_1} \log\left(1 - \lambda + \lambda e^{\alpha+\beta y_j}\right) + 2\log(\lambda), \end{aligned} \tag{2.6}$$

where ξ is the solution to (2.5). The penalty function $\log(\lambda)$ goes to $-\infty$ as λ approaches 0. Therefore, λ is bounded away from 0, and the null hypothesis in (2.3) then reduces to $(\alpha, \beta) = (0, 0)$. That is, the parameters in the penalized log-EL function is asymptotically identifiable. However, the asymptotic behavior of the penalized ELR test is still complicated. Meanwhile, the computation of the penalized ELR test statistic is another obstacle of the implementation of the penalized ELR method. No feasible and stable algorithm has been found for this purpose. An EL-based EM test proposed in this paper provides an efficient way to solve the problem.

3. EL-Based EM Test

Motivated by Chen and Li [12] and Li et al. [13], we propose an EM test based on the penalized EL to test the hypothesis (2.3). The EM test statistics are derived iteratively. We first choose a finite set of $\Lambda = \{\lambda_1, \ldots, \lambda_L\} \subset (0,\ 1]$, for instance, $\Lambda = \{0.1, 0.2, \ldots, 0.9, 1.0\}$, and a positive integer K (2 or 3 in general). For each $l = 1, \ldots, L$, we proceed the following steps.

Step 1. Let $k = 1$ and $\lambda_l^{(k)} = \lambda_l$. Calculate $(\alpha_l^{(k)}, \beta_l^{(k\ )}) = \text{argmax}_{\alpha,\beta}\ p\ R(\lambda_l^{(k)}\ , \alpha,\ \beta)$.

Step 2. Update (λ, α, β) by using the following algorithm for $K - 1$ times.

Substep 2.1. Calculate the posterior distribution,

$$w_{jl}^{(k)} = \frac{\lambda_l^{(k)}\ \exp\left(\alpha_l^{(k)} + \beta_l^{(k)} y_j\right)}{1 - \lambda_l^{(k)} + \lambda_l^{(k)} \exp\left(\alpha_l^{(k)} + \beta_l^{(k)} y_j\right)}, \quad j = 1, \ldots, n_1, \tag{3.1}$$

and update λ by

$$\lambda_l^{(k+1)} = \arg\max_{\lambda} \left\{ \sum_{j=1}^{n_1} \left(1 - w_{jl}^{(k)}\right) \log(1 - \lambda) + \sum_{j=1}^{n_1} w_{jl}^{(k)} \log(\lambda) + \log(\lambda) \right\}. \tag{3.2}$$

Substep 2.2. Update (α, β) by $(\alpha_l^{(k+1)}, \beta_l^{(k+1)}) = \text{argmax}_{\alpha,\beta} pR(\lambda_l^{(k+1)}, \alpha, \beta)$.

Substep 2.3. Let $k = k + 1$ and continue.

Step 3. Define the test statistics $M_n^{(K)}(\lambda_l) = pR(\lambda_l^{(K)}, \alpha_l^{(K)}, \beta_l^{(K)})$.

The EM test statistic is defined as

$$\text{EM}_n^{(K)} = \max\left\{ M_n^{(K)}(\lambda_l), l = 1, \ldots, L \right\}. \tag{3.3}$$

We reject the null hypothesis H_0 when the EM test statistic is greater than some critical value determined by the following limiting distribution.

Theorem 3.1. *Suppose* $\rho = n_1/n \in (0, 1)$ *is a constant. Assume the null hypothesis* H_0 *holds and* $E(t_h) = 0$ *and* $\text{Var}(t_h)\ = \sigma^2 \in (0,\infty)$ *for* $h = 1, \ldots, n$*. For* $l = 1, \ldots, L$ *and any fixed* k*, it holds that*

$$\lambda_l^{(k)} - \lambda_l = o_p(1), \qquad \alpha_l^{(k)} = O_p\left(n^{-1}\right), \qquad \beta_l^{(k)} = \frac{\overline{y} - \overline{x}}{\lambda_l \sigma^2} + o_p\left(n^{-1/2}\right), \tag{3.4}$$

where $\overline{x} = (1/n_0) \sum_{i=1}^{n_0} x_i$ *and* $\overline{y} = (1/n_1) \sum_{j=1}^{n_1} y_j$*.*

Remark 3.2. The assumption $Et_h = 0$ is only for convenience purpose and unnecessary. Otherwise, we can replace t_h and α with $t_h - E(t_h)$ *and* $\alpha + \beta E(t_h)$.

Theorem 3.3. *Assume the conditions of Theorem 3.1 hold and* $1 \in \Lambda$*. Under the null hypothesis* (2.3), $\text{EM}_n^{(K)} \rightarrow \varnothing_1^2$ *in distribution, as* $n \rightarrow \infty$*.*

We finish this section with an additional remark.

Remark 3.4. We point out that the idea of the EM-test can also be generalized to more general models such as $\log(g(x)/f(x)) = \alpha + \beta_1 x + \cdots + \beta_k x^k$ for some integer k or $\log(g(x)/f(x)) = \alpha + \beta_1 t_1(x) + \cdots + \beta_k t_k(x)$ with $t_i(\cdot)$'s being known functions.

4. Computation of the EM Test

A key step of the EM test procedure is to maximize $pR(\lambda, \alpha, \beta)$ with respect to (α, β) for fixed λ. In this section, we propose a computation strategy which provides stable solution to this optimization problem. Throughout this section, λ is suppressed to be fixed.

The objective function is $pR(\lambda, \alpha, \beta) = G(\xi_*, \alpha, \beta)$ where

$$G(\xi, \alpha, \beta) = -2\sum_{h=1}^{n} \log\left[1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)\right] + 2\sum_{j=1}^{n_1} \log\left[1 - \lambda + \lambda e^{\alpha+\beta y_j}\right] + 2\log(\lambda) \tag{4.1}$$

and $\xi_* = \xi_*(\alpha, \beta)$ is the solution to

$$\frac{\partial G}{\partial \xi} = -2\sum_{h=1}^{n} \frac{e^{\alpha+\beta t_h} - 1}{1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)} = 0. \tag{4.2}$$

If (α, β) is the maximum point of $pR(\lambda, \alpha, \beta)$, it should generally satisfy

$$\frac{\partial G}{\partial \alpha} = -2\sum_{h=1}^{n} \frac{\xi e^{\alpha+\beta t_h}}{1 + \xi\left(e^{\alpha+\beta t_h} - 1\right)} + 2\sum_{j=1}^{n_1} \frac{\lambda e^{\alpha+\beta y_j}}{1 - \lambda + \lambda e^{\alpha+\beta y_j}} = 0. \tag{4.3}$$

Combining (4.2) and (4.3) leads to

$$\xi = \frac{1}{n}\sum_{j=1}^{n_1} \frac{\lambda e^{\alpha+\beta y_j}}{1 - \lambda + \lambda e^{\alpha+\beta y_j}}. \tag{4.4}$$

Putting this expression of ξ back into (4.1), we have a new function

$$\begin{aligned} H(\alpha, \beta) = {} & -2\sum_{h=1}^{n} \log\left\{1 + \left(e^{\alpha+\beta t_h} - 1\right)\frac{1}{n}\sum_{j=1}^{n_1} \frac{\lambda e^{\alpha+\beta y_j}}{1 - \lambda + \lambda e^{\alpha+\beta y_j}}\right\} \\ & + 2\sum_{j=1}^{n_1} \log\left(1 - \lambda + \lambda e^{\alpha+\beta y_j}\right). \end{aligned} \tag{4.5}$$

It can be verified that $H(\alpha, \beta)$ is almost surely concave in a neighborhood of $(0, 0)$ given λ, which means that maximizing $H(\alpha, \beta)$ with respect to (α, β) gives the maximum of $pR(\lambda, \alpha, \beta)$ for fixed λ. The stability of the method is illustrated by the following simulation study.

5. Simulation Study

We consider two models in Examples 1.3 and 1.4 with $\mu_1 = 0$, $\mu_2 = \mu$, and $\sigma^2 = 1$ for Example 1.3, and $m_1 = 1$, $m_2 = m$, and $\theta = 1$ for Example 1.4. Nominal levels of 0.01, 0.05, and 0.10 are considered. The logarithm transformation is applied to the original data before using the EM test. The initial set $\Lambda = \{0.1, 0.2, \ldots, 1\}$ and iteration number $K = 3$ are used to calculate the EM test statistic.

One competitive method for testing homogeneity under the semiparametric two-sample model is the score test proposed by Qin and Liang [1]. This method is based on

$$S(\alpha, \beta) = \left.\frac{\partial l(\lambda, \alpha, \beta)}{\partial \lambda}\right|_{\lambda=0} = \sum_{j=1}^{n_1}\left(e^{\alpha+\beta y_j} - 1\right), \tag{5.1}$$

where $l(\lambda, \alpha, \beta)$ is the log empirical likelihood function given in (2.4). Let $(\widehat{\alpha}_1, \widehat{\beta}_1) = \text{argmax}_{\alpha,\beta} l(1, \alpha, \beta)$. The score test statistic was defined as $T_1 = S(\widehat{\alpha}_1, \widehat{\beta}_1)/(1+n_1/n_0)$, which has a χ_1^2 limiting distribution under the null hypothesis.

We compare the EM test and the score test in terms of type I error and power. We calculate the type I errors of each method under the null hypothesis based on 20,000 repetitions and the power under the alternative models based on 2,000 repetitions. For fair comparison, simulated critical values are used to calculate the power. We consider two sample sizes: 50 and 200 and $K = 1, 2, 3$. Tables 1 and 2 contain the simulation results for the log-normal models and Tables 3 and 4 for the gamma models.

The results show that the EM test and the score test have similar type I errors. For both methods, the type I errors are somehow larger than the nominal levels when the sample size is $n = 50$; they are close to the nominal levels when the sample size is increased to $n = 200$. For the log-normal models, two methods have almost the same power when the alternatives are close to each other such as $\mu = 1$; the EM test becomes much more powerful when the alternatives are distant and the sample size increases. In the case of $n = 50$, $\lambda = 0.2$, $\mu = 3$, and nominal level 0.01, the EM test has a 10% gain in power compared with the score test; the gain rushes up to almost 30% when $\lambda = 0.1$, $\mu = 3$, and the sample size increases to $n = 200$. For the gamma models, the advantage of the EM test is more obvious. For both sample sizes $n = 50$ and 200, the EM test is more powerful than the score test.

6. Real Example

We apply our EM test procedure to the drug abuse data [15] in a study of addiction to morphine in rats. In this study, rats got morphine by pressing a lever and the frequency of lever presses (self-injection rates) after six-day treatment with morphine was recorded as response variable. The data consist of the number of lever presses for five groups of rats: four treatment groups with different dose levels and one saline group (control group).

We analyzed the response variables (the number of lever presses by rats) of the treatment group at the first dose level and the control group. The data is tabulated in Table 3 of Fu et al. [5]. Following Boos and Browine [16] and Fu et al. [5], we analyze the transformed data, $\log_{10}(R+1)$ with R being the number of lever presses by rats. Instead of using the parametric models as Boos and Browine [16] and Fu et al. [5], we adopt Anderson's semiparametric approach. That is, we assume that the response variables in control group comes from $f(x)$,

Table 1: Type I error and power comparisons (%) of the EM test and the score test (SC test) for log-normal model: $n_0 = n_1 = 50$.

λ	μ	Level	$\mathrm{EM}_n^{(1)}$	$\mathrm{EM}_n^{(2)}$	$\mathrm{EM}_n^{(3)}$	SC test
0		10	11.9	12.2	12.2	11.5
0		5	6.3	6.5	6.5	6.4
0		1	1.6	1.6	1.6	1.9
0.1	1	10	14.8	14.5	14.5	14.6
0.1	1	5	8.5	8.6	8.6	8.6
0.1	1	1	2.5	2.5	2.5	2.4
0.1	2	10	27.2	28.1	28	25.6
0.1	2	5	17.8	18.4	18.4	16.7
0.1	2	1	6.6	6.7	6.7	6.2
0.1	3	10	47.1	48.3	48.3	41.4
0.1	3	5	34.2	35.6	35.4	30.8
0.1	3	1	15.4	15.6	15.6	14.9
0.2	1	10	25.5	25.9	26	25.6
0.2	1	5	16.4	16.4	16.4	17
0.2	1	1	5.4	5.3	5.3	5.7
0.2	2	10	62.2	62.7	62.7	56.7
0.2	2	5	50.6	51.3	51.2	45.9
0.2	2	1	28.4	28.5	28.5	24.7
0.2	3	10	88.3	88.8	88.8	81
0.2	3	5	81	82.3	82.3	73.4
0.2	3	1	61.7	61.9	61.9	51.5
0.3	1	10	43.3	42.9	42.8	42.8
0.3	1	5	31.3	31.1	31.1	31.6
0.3	1	1	14.2	14.2	14.2	13.9
0.3	2	10	88.1	88.5	88.5	84.2
0.3	2	5	80.8	80.8	80.8	76.8
0.3	2	1	61.5	61.5	61.5	55.3
0.3	3	10	99.3	99.3	99.3	97
0.3	3	5	98	98.2	98.2	94.8
0.3	3	1	93	93.2	93.2	85.2

while the response variables in treatment group comes from $h(x) = (1-\lambda)f(x) + \lambda g(x)$ with $g(x)/f(x) = \exp(\alpha + \beta x)$. The EM test statistics for testing homogeneity under the

Table 2: Type I error and power comparisons (%) of the EM test and the score test (SC test) for log-normal model: $n_0 = n_1 = 200$.

λ	μ	Level	$\mathrm{EM}_n^{(1)}$	$\mathrm{EM}_n^{(2)}$	$\mathrm{EM}_n^{(3)}$	SC test
0		10	10.4	10.5	10.6	10.2
0		5	5.5	5.6	5.6	5.4
0		1	1.2	1.2	1.2	1.2
0.1	1	10	26.5	26.7	26.5	26.2
0.1	1	5	17.2	17.2	17.2	16.4
0.1	1	1	5.8	5.9	6	5.6
0.1	2	10	68.3	69	69.2	58.4
0.1	2	5	58.5	58.8	58.9	47.4
0.1	2	1	37	37.1	37.4	25.1
0.1	3	10	96.4	96.8	97	84.4
0.1	3	5	94.6	94.8	95.2	77.6
0.1	3	1	86.2	87.2	87.4	58.6
0.2	1	10	63	62.9	62.8	62.1
0.2	1	5	50.2	50	50	49.4
0.2	1	1	27.8	27.6	27.5	26.2
0.2	2	10	99.2	99.3	99.4	97.5
0.2	2	5	98.6	98.6	98.6	95
0.2	2	1	95.1	95.2	95.2	85.5
0.2	3	10	100	100	100	100
0.2	3	5	100	100	100	99.9
0.2	3	1	100	100	100	99.2
0.3	1	10	89.5	89.5	89.6	89
0.3	1	5	84	83.9	83.9	82.6
0.3	1	1	65.1	64.9	64.6	63
0.3	2	10	100	100	100	100
0.3	2	5	100	100	100	99.9
0.3	2	1	100	100	100	99.7
0.3	3	10	100	100	100	100
0.3	3	5	100	100	100	100
0.3	3	1	100	100	100	100

semiparametric two-sample model are found to be $\mathrm{EM}_n^{(1)} = 14.090$, $\mathrm{EM}_n^{(2)} = 14.150$, and $\mathrm{EM}_n^{(3)} = 14.167$. Calibrated by the χ_1^2 limiting distribution, the P values are all around 0.02%.

Table 3: Type I error and power comparisons (%) of the EM test and the score test (SC test) for gamma model: $n_0 = n_1 = 50$.

λ	m	Level	$\mathrm{EM}_n^{(1)}$	$\mathrm{EM}_n^{(2)}$	$\mathrm{EM}_n^{(3)}$	SC test
0		10	12.2	12.5	12.5	12.1
0		5	6.4	6.6	6.6	6.7
0		1	1.4	1.4	1.4	2.3
0.1	2	10	14.9	15.1	15.2	12
0.1	2	5	8.8	8.9	8.9	6.4
0.1	2	1	2.8	2.8	2.8	0.6
0.1	3	10	19.6	19.9	19.9	14.1
0.1	3	5	13.2	13.2	13.2	7.7
0.1	3	1	4.3	4.4	4.4	1
0.1	4	10	25.5	26.4	26.5	17
0.1	4	5	17.5	17.9	17.9	9.2
0.1	4	1	6.3	6.4	6.4	1.1
0.2	2	10	22.9	22.7	22.8	17.6
0.2	2	5	14.4	14.3	14.3	9.2
0.2	2	1	4.5	4.7	4.7	1.2
0.2	3	10	39.6	39.9	40	27.4
0.2	3	5	29.1	29.5	29.5	16.7
0.2	3	1	14.3	14.4	14.4	4
0.2	4	10	61.1	61.7	61.7	37
0.2	4	5	49.2	49.6	49.6	24.1
0.2	4	1	28.4	28.6	28.6	6.6
0.3	2	10	36.3	36.4	36.4	28.6
0.3	2	5	26.1	25.9	25.9	16.9
0.3	2	1	11.9	11.9	11.9	3.1
0.3	3	10	67.2	67.2	67.2	48.9
0.3	3	5	55.8	55.8	55.8	35.1
0.3	3	1	34	34	34.1	11.3
0.3	4	10	87.9	88.1	88.2	67.5
0.3	4	5	81.8	82.2	82.2	53.4
0.3	4	1	63.1	63.3	63.4	21.4

We also applied the score test of Qin and Liang [1]. The score test statistic is 9.417 with the P value equal to 0.2% calibrated by the χ_1^2 limiting distribution. We also used the permutation

Table 4: Type I error and power comparisons (%) of the EM test and the score test (SC test) for gamma model: $n_0 = n_1 = 200$.

λ	m	Level	$\mathrm{EM}_n^{(1)}$	$\mathrm{EM}_n^{(2)}$	$\mathrm{EM}_n^{(3)}$	SC test
0		10	11.2	11.3	11.3	11.1
0		5	5.9	5.9	6	5.7
0		1	1.2	1.2	1.2	1.4
0.1	2	10	23.1	22.7	22.7	19.7
0.1	2	5	14.2	14.2	14.2	11.7
0.1	2	1	5.1	5.1	5.2	3.1
0.1	3	10	39.6	39.8	39.9	29.5
0.1	3	5	29	29.4	29.6	19
0.1	3	1	13.2	13.4	13.5	4.8
0.1	4	10	62.3	62.5	62.7	37.5
0.1	4	5	52.2	52.5	52.8	26.2
0.1	4	1	32.5	33.2	33.7	8.5
0.2	2	10	49	48.9	48.9	43.8
0.2	2	5	36.6	36.7	36.5	30.6
0.2	2	1	19.4	19.4	19.4	11.4
0.2	3	10	88.2	88.2	88.4	73
0.2	3	5	81.5	81.6	81.6	61.2
0.2	3	1	64.6	64.6	64.8	34.6
0.2	4	10	98.9	98.9	98.9	87.1
0.2	4	5	98	98.1	98.1	79.7
0.2	4	1	94.3	94.2	94.2	54.5
0.3	2	10	78.5	78.5	78.6	73
0.3	2	5	70.1	70	70	62.5
0.3	2	1	48.7	48.8	48.8	34.9
0.3	3	10	99.2	99.2	99.2	96.1
0.3	3	5	98.8	98.8	98.8	93
0.3	3	1	96.5	96.5	96.5	78.8
0.3	4	10	100	100	100	99.4
0.3	4	5	100	100	100	98.7
0.3	4	1	100	100	100	92.5

methods to get the P values of the two types of tests. Based on 50,000 permutations, the P values of the three EM test statistics are all around 0.03%, and the P value of the score test is around 0.5%. In accordance with Fu et al. [5], both methods suggest a significant treatment effect, while the proposed EM test has much stronger evidence than the score test.

Appendix

Proofs

The proofs of Theorems 3.1 and 3.3 are based on the three lemmas given below. Lemma A.1 assesses the order of the maximum empirical likelihood estimators of α and β with λ bounded away from 0 under the null hypothesis. Lemma A.2 shows that the EM iteration updates the value of λ by the amount of order $o_p(1)$. Theorem 3.1 is then proved by iteratively using Lemmas A.1 and A.2. Lemma A.3 gives an approximation of the penalized ELR for any λ bounded away from 0, based on which we prove Theorem 3.3.

Lemma A.1. *Assume the conditions of Theorem 3.1. Let* $\overline{\lambda} \in [\epsilon, 1]$ *for some constant* $\epsilon > 0$ *and* $(\overline{\alpha}, \overline{\beta}) = \operatorname{argmax}_{\alpha,\beta} pR(\overline{\lambda}, \alpha, \beta)$. *Then, we have*

$$\overline{\alpha} = O_p\left(n^{-1}\right), \quad \overline{\beta} = \frac{\overline{y} - \overline{x}}{\overline{\lambda}\sigma^2} + o_p\left(n^{-1/2}\right) \tag{A.1}$$

with $\overline{x} = 1/n_0 \sum_{i=1}^{n_0} x_i$ *and* $\overline{y} = 1/n_1 \sum_{j=1}^{n_1} y_j$.

Proof. Since $\overline{\lambda} \geq \epsilon > 0$, the parameters (α, β) in the empirical likelihood ratio are identifiable. Therefore, $(\overline{\alpha}, \overline{\beta})$ are $\sqrt{n}$-consistent to the true value $(0, 0)$, that is, $\overline{\alpha} = O_p(n^{-1/2})$ and $\overline{\beta} = O_p(n^{-1/2})$ [10].

Following the arguments in Section 4, the maximum empirical likelihood estimate $(\overline{\alpha}, \overline{\beta})$ should satisfy (here λ is suppressed to $\overline{\lambda}$)

$$\frac{\partial G\left(\overline{\xi}, \overline{\alpha}, \overline{\beta}\right)}{\partial \alpha} = -2\sum_{h=1}^{n} \frac{\overline{\xi} e^{\overline{\alpha}+\overline{\beta}t_h}}{1 + \overline{\xi}\left(e^{\overline{\alpha}+\overline{\beta}t_h} - 1\right)} + 2\sum_{j=1}^{n_1} \frac{\overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}}{1 - \overline{\lambda} + \overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}} = 0, \tag{A.2}$$

$$\frac{\partial G\left(\overline{\xi}, \overline{\alpha}, \overline{\beta}\right)}{\partial \beta} = -2\sum_{h=1}^{n} \frac{\overline{\xi} e^{\overline{\alpha}+\overline{\beta}t_h} t_h}{1 + \overline{\xi}\left(e^{\overline{\alpha}+\overline{\beta}t_h} - 1\right)} + 2\sum_{j=1}^{n_1} \frac{\overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j} y_j}{1 - \overline{\lambda} + \overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}} = 0 \tag{A.3}$$

with

$$\overline{\xi} = \frac{1}{n}\sum_{j=1}^{n_1} \frac{\overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}}{1 - \overline{\lambda} + \overline{\lambda} e^{\overline{\alpha}+\overline{\beta}y_j}}. \tag{A.4}$$

Applying Taylor expansion on the right-hand side of (A.4), we get

$$\overline{\xi} = \frac{n_1}{n}\overline{\lambda} + o_p(1). \tag{A.5}$$

Further applying first-order Taylor expansion to (A.2) and using (A.4), we get

$$n\overline{\xi}\left(1-\overline{\xi}\right)\overline{\alpha}+\overline{\xi}\left(1-\overline{\xi}\right)\sum_{h=1}^{n}t_h\overline{\beta}-n_1\overline{\lambda}\left(1-\overline{\lambda}\right)\overline{\alpha}-\overline{\lambda}\left(1-\overline{\lambda}\right)\sum_{j=1}^{n_1}y_j\overline{\beta}=O_p(n)\left(\overline{\alpha}^2+\overline{\beta}^2\right). \tag{A.6}$$

Note that both $\overline{\alpha}$ and $\overline{\beta}$ are of order $O_p(n^{-1/2})$ and that both $\sum_{h=1}^{n} t_h$ and $\sum_{j=1}^{n_1} y_j$ have order $O_p(n^{1/2})$. Combining (A.5) and (A.6) yields $n_1\overline{\lambda}(\overline{\lambda}-n_1/n\overline{\lambda})\overline{\alpha}=O_p(1)$. Therefore, $\overline{\alpha}=O_p(n^{-1})$.

Similarly, first-order Taylor expansion of (A.3) results in

$$\begin{aligned}0=&-\overline{\xi}\sum_{h=1}^{n}t_h-\overline{\xi}\left(1-\overline{\xi}\right)\sum_{h=1}^{n}t_h\overline{\alpha}-\overline{\xi}\left(1-\overline{\xi}\right)\sum_{h=1}^{n}t_h^2\overline{\beta}\\&+\overline{\lambda}\sum_{j=1}^{n_1}y_j+\overline{\lambda}\left(1-\overline{\lambda}\right)\sum_{j=1}^{n_1}y_j\overline{\alpha}+\ \overline{\lambda}\left(1-\overline{\lambda}\right)\sum_{j=1}^{n_1}y_j^2\overline{\beta}+O_p(n)\left(\overline{\alpha}^2+\overline{\beta}^2\right).\end{aligned} \tag{A.7}$$

With the same reasoning as for $\overline{\alpha}$, it follows from (A.7) that

$$\left\{n_1\overline{\lambda}\left(1-\frac{n_1}{n\overline{\lambda}}\right)\sigma^2-n_1\overline{\lambda}\left(1-\overline{\lambda}\right)\sigma^2\right\}\overline{\beta}=\overline{\lambda}\sum_{j=1}^{n_1}y_j-\frac{n_1}{n}\overline{\lambda}\sum_{h=1}^{n}t_h+o_p\left(n^{1/2}\right). \tag{A.8}$$

After some algebra, we have $\overline{\beta}=(\overline{y}-\overline{x})/(\overline{\lambda}\sigma^2)+o_p(n^{-1/2})$, which completes the proof. □

Suppose that $\overline{\lambda}$, $\overline{\alpha}$, and $\overline{\beta}$ have the properties given in Lemma A.1. For $j=1,\ldots,n_1$, let $\overline{w}_j=\overline{\lambda}\exp(\overline{\alpha}+\overline{\beta}y_j)/(1-\overline{\lambda}+\overline{\lambda}\exp(\overline{\alpha}+\overline{\beta}y_j))$. The updated value of λ is

$$\overline{\lambda}^*=\arg\max_{\lambda}\left\{\sum_{j=1}^{n_1}(1-\overline{w}_j)\log(1-\lambda)+\sum_{j=1}^{n_1}\overline{w}_j\log(\lambda)+\log(\lambda)\right\}. \tag{A.9}$$

It can be verified that the close form of $\overline{\lambda}^*$ is given by $\overline{\lambda}^*=(1/(n_1+1))(\sum_{j=1}^{n_1}\overline{w}_j+1)$. We now show that the above iteration only changes the value of λ by an $o_p(1)$ term.

Lemma A.2. *Assume the conditions of Lemma A.1 hold. Then,* $\overline{\lambda}^*=\overline{\lambda}+o_p(1)$.

Proof. Let $\widehat{\lambda}=\sum_{j=1}^{n_1}\overline{w}_j/n_1$. According to Lemma A.1, $\overline{\alpha}=o_p(1)$ and $\overline{\beta}=o_p(1)$. Applying the first-order Taylor expansion, we have

$$\widehat{\lambda}=\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\overline{\lambda}\exp\left(\overline{\alpha}+\overline{\beta}y_j\right)}{1-\overline{\lambda}+\overline{\lambda}\exp\left(\overline{\alpha}+\overline{\beta}y_j\right)}=\overline{\lambda}+O_p(1)\ \left(\overline{\alpha}+\overline{\beta}\right)=\overline{\lambda}+o_p(1). \tag{A.10}$$

Some simple algebra work shows that

$$\overline{\lambda}^* - \widehat{\lambda} = \frac{1-\overline{\lambda}}{n_1+1} = o_p(1). \tag{A.11}$$

Therefore, $\overline{\lambda}^* = \overline{\lambda} + o_p(1)$, and this finishes the proof. □

Proof of Theorem 3.1. With the above two technical lemmas, the proof is the same as that of Theorem 1 in Li et al. [13] and therefore is omitted. □

The next lemma is a technical preparation for proving Theorem 3.3. It investigates the asymptotic approximation of the penalized ELR for any λ bounded away from 0.

Lemma A.3. *Assume the conditions of Theorem 3.1 and $\overline{\lambda} \in [\epsilon, 1]$ for some $\epsilon > 0$. Then,*

$$pR\left(\overline{\lambda},\overline{\alpha},\overline{\beta}\right) = n\rho(1-\rho)\sigma^{-2}(\overline{y}-\overline{x})^2 + 2\log\left(\overline{\lambda}\right) + o_p(1). \tag{A.12}$$

Proof. With Lemma A.1, we have $\overline{\alpha} = O_p(n^{-1})$ and $\overline{\beta} = O_p(n^{-1/2})$. Applying second-order Taylor expansion on $pR(\overline{\lambda},\overline{\alpha},\overline{\beta})$ and noting that $\partial pR/\partial\alpha|_{(\alpha,\beta)=(0,0)} = 0$, we have

$$\begin{aligned} pR\left(\overline{\lambda},\overline{\alpha},\overline{\beta}\right) = {} & 2\left(-\overline{\xi}\sum_{h=1}^{n} t_h + \overline{\lambda}\sum_{j=1}^{n_1} y_j\right)\overline{\beta} - \left\{\overline{\xi}\left(1-\overline{\xi}\right)\sum_{h=1}^{n} t_h^2 - \overline{\lambda}\left(1-\overline{\lambda}\right)\sum_{j=1}^{n_1} y_j^2\right\}\overline{\beta}^2 \\ & + 2\log\left(\overline{\lambda}\right) + o_p(1). \end{aligned} \tag{A.13}$$

Using (A.5) and the facts that both $\sum_{h=1}^{n} t_h^2/n$ and $\sum_{j=1}^{n_1} y_j^2/n_1$ converge to σ^2 in probability, the above expression can be simplified to

$$pR\left(\overline{\lambda},\overline{\alpha},\overline{\beta}\right) = 2\frac{n_1 n_0}{n}\overline{\lambda}(\overline{y}-\overline{x})\overline{\beta} - \frac{n_1 n_0}{n}\overline{\lambda}^2\sigma^2\overline{\beta}^2 + 2\log\left(\overline{\lambda}\right) + o_p(1). \tag{A.14}$$

Plugging in the approximation $\overline{\beta} = (\overline{y}-\overline{x})/(\overline{\lambda}\sigma^2) + o_p(n^{-1/2})$, we get

$$\begin{aligned} pR\left(\overline{\lambda},\overline{\alpha},\overline{\beta}\right) &= \frac{n_1 n_0}{n}\frac{(\overline{y}-\overline{x})^2}{\sigma^2} + 2\log\left(\overline{\lambda}\right) + o_p(1) \\ &= n\rho(1-\rho)\sigma^{-2}(\overline{y}-\overline{x})^2 + 2\log\left(\overline{\lambda}\right) + o_p(1). \end{aligned} \tag{A.15}$$

This completes the proof. □

Proof of Theorem 3.3. Without loss of generality, we assume $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_L = 1$. According to Theorem 3.1 and Lemma A.3, for $l = 1, \ldots, L$, we have

$$pR\left(\lambda_l^{(K)},\alpha_l^{(K)},\beta_l^{(K)}\right) = n\rho(1-\rho)\sigma^{-2}(\overline{y}-\overline{x})^2 + 2\log(\lambda_l) + o_p(1). \tag{A.16}$$

This leads to

$$\mathrm{EM}_n^{(K)} = \max_{1\leq l\leq L} pR\left(\lambda_l^{(K)}, \alpha_l^{(K)}, \beta_l^{(K)}\right) = n\rho(1-\rho)\sigma^{-2}(\overline{y}-\overline{x})^2 + o_p(1), \tag{A.17}$$

where the remainder is still $o_p(1)$ since the maximum is taken over a finite set.

Note that when n tends to infinity, $\sqrt{n}(\overline{y}-\overline{x}) \longrightarrow N(0,\ \sigma^2/[\rho(1-\rho)])$ in distribution. Therefore,

$$\mathrm{EM}_n^{(K)} \longrightarrow \chi_1^2 \tag{A.18}$$

in distribution as n goes to infinity. This completes the proof. □

References

[1] J. Qin and K. Y. Liang, "Hypothesis testing in a mixture case-control model," *Biometrics*, vol. 67, pp. 182–193, 2011.

[2] J. Zhang, "Powerful two-sample tests based on the likelihood ratio," *Technometrics*, vol. 48, no. 1, pp. 95–103, 2006.

[3] J. A. Anderson, "Multivariate logistic compounds," *Biometrika*, vol. 66, no. 1, pp. 17–26, 1979.

[4] T. Lancaster and G. Imbens, "Case-control studies with contaminated controls," *Journal of Econometrics*, vol. 71, no. 1-2, pp. 145–160, 1996.

[5] Y. Fu, J. Chen, and J. D. Kalbfleisch, "Modified likelihood ratio test for homogeneity in a two-sample problem," *Statistica Sinica*, vol. 19, no. 4, pp. 1603–1619, 2009.

[6] A. B. Owen, "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, vol. 75, no. 2, pp. 237–249, 1988.

[7] A. B. Owen, "Empirical likelihood ratio confidence regions," *The Annals of Statistics*, vol. 18, no. 1, pp. 90–120, 1990.

[8] P. Hall and B. La Scala, "Methodology and algorithms of empirical likelihood," *International Statistical Review*, vol. 58, pp. 109–127, 1990.

[9] T. DiCiccio, P. Hall, and J. Romano, "Empirical likelihood is Bartlett-correctable," *The Annals of Statistics*, vol. 19, no. 2, pp. 1053–1061, 1991.

[10] J. Qin and J. Lawless, "Empirical likelihood and general estimating equations," *The Annals of Statistics*, vol. 22, no. 1, pp. 300–325, 1994.

[11] S. E. Ahmed, A. Hussein, and S. Nkurunziza, "Robust inference strategy in the presence of measurement error," *Statistics & Probability Letters*, vol. 80, no. 7-8, pp. 726–732, 2010.

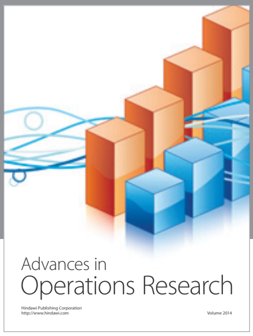
[12] J. Chen and P. Li, "Hypothesis test for normal mixture models: the EM approach," *The Annals of Statistics*, vol. 37, no. 5, pp. 2523–2542, 2009.

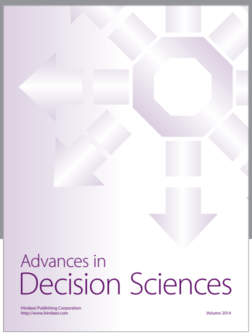
[13] P. Li, J. Chen, and P. Marriott, "Non-finite Fisher information and homogeneity: an EM approach," *Biometrika*, vol. 96, no. 2, pp. 411–426, 2009.

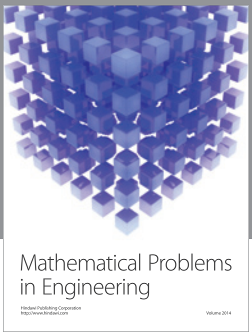
[14] J. Chen, "Penalized likelihood-ratio test for finite mixture models with multinomial observations," *The Canadian Journal of Statistics*, vol. 26, no. 4, pp. 583–599, 1998.

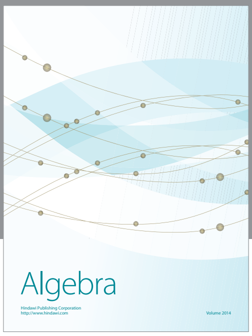
[15] J. R. Weeks and R. J. Collins, "Primary addiction to morphine in rats," *Federation Proceedings*, vol. 30, p. 277, 1971.

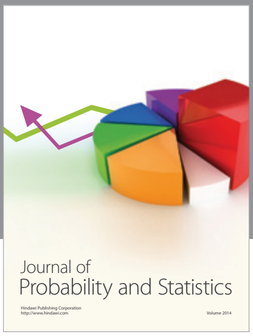
[16] D. D. Boos and C. Brownie, "Mixture models for continuous data in dose-response studies when some animals are unaffected by treatment," *Biometrics*, vol. 47, pp. 1489–1504, 1991.


Advances in
Operations Research

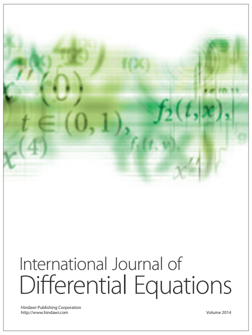

Advances in
Decision Sciences

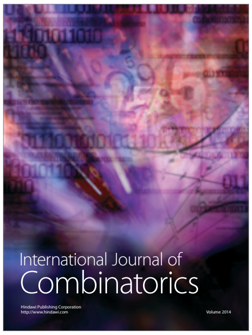

Mathematical Problems
in Engineering

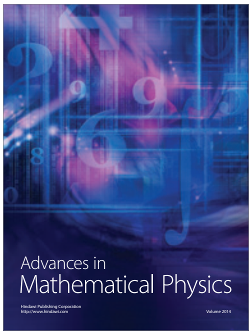

Algebra

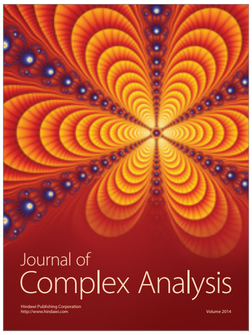

Journal of
Probability and Statistics

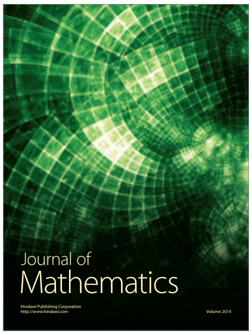
The Scientific
World Journal

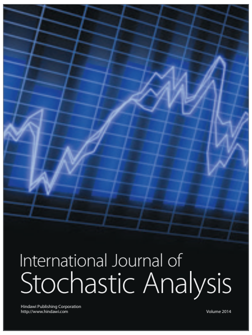

International Journal of
Differential Equations

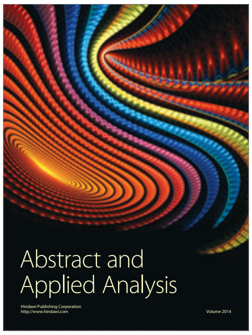

International Journal of
Combinatorics

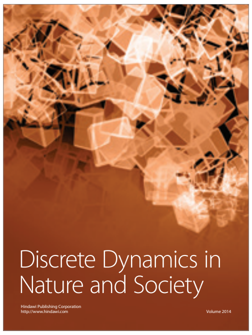

Advances in
Mathematical Physics

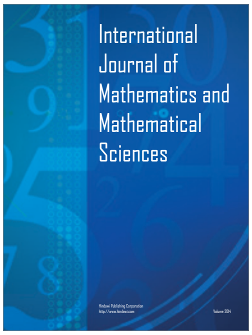

Journal of
Complex Analysis


Journal of
Mathematics


International Journal of
Stochastic Analysis

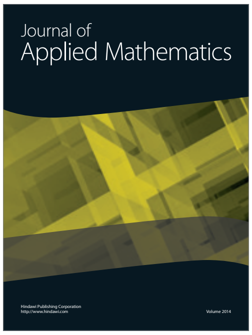

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


Journal of
Applied Mathematics

Journal of
Optimization