Malaysian Journal of Science **26** (SKSM14 Special Issue): 139 - 143 (2007)

Error Exponents in Hypothesis Testing and Chernoff Information

Choon Peng Tan

Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, 13, Jalan 13/6,
46200 Petaling Jaya, Selangor, Malaysia
tancp@mail.utar.edu.my

ABSTRACT We consider a simple hypothesis testing problem on the parameters of a probability distribution belonging to the exponential class. It is well-known that the Chernoff information is the best asymptotic achievable exponent in the Bayesian probability of error when we use a likelihood ratio test with an exponential threshold function of the sample size. We shall derive the general forms of the error exponent and the Chernoff information for the exponential class. In tests using the maximum-a-posteriori probability decision rule, the Chernoff information provides a lower bound on the error exponent. The Chernoff informations of some common distributions will be demonstrated.

ABSTRAK Kami mempertimbangkan satu masalah pengujian hipotesis mudah mengenai parameter suatu taburan kebarangkalian yang datang dari kelas eksponen. Maklumat Chernoff adalah terkenal sebagai eksponen bolehcapai asimptot terbaik dalam ralat kebarangkalian Bayesan bila ujian nisbah kebolehjadian digunakan dengan fungsi treshold eksponen dalam saiz sampel. Kami akan terbitkan bentuk am bagi eksponen ralat dan maklumat Chernoff untuk kelas eksponen. Dalam ujian yang menggunakan petua keputusan kebarangkalian-posterior-maksimum, maklumat Chernoff memberi satu batas bawah untuk eksponen ralat. Maklumat Chernoff bagi sesetengah taburan biasa yang telah ditemui akan ditunjukkan.

(Hypothesis testing, error exponent, Chernoff information, exponential class)

INTRODUCTION

It is well-known that in a fixed-sample-size two-hypotheses testing problem, the probability of the Type I error can be controlled. The probability of the Type II error goes to zero exponentially fast with a best rate given by Stein's Lemma [1], as the sample size n increases to infinity. Using Sanov's Theorem [1], [2], the error exponents of both types of probabilities of error can be determined, assuming that an optimal test using the Neyman-Pearson Lemma is used. These exponents are relative entropies of certain probability functions or probability density functions associated with the test. For the symmetric case where both exponents are equal, this exponent is known as the Chernoff information. The Chernoff information is the best error exponent in likelihood ratio tests with an exponential threshold function of the sample size [1]. In maximum-a-posteriori-probability Bayesian tests, the best error exponent is bounded below by the Chernoff information. Hence the Chernoff information provides an estimate of the best error exponent in such tests.

Let $\{f(x;\theta):\theta\in\Theta\}$ be a parametric family of probability density functions of X if X is continuous or a parametric family of probability mass functions of X if X is discrete. Without loss of generality, we assume that X is continuous. The family $\{f(x;\theta):\theta\in\Theta\}$ is said to be from the *multi-parameter or vector-valued exponential class* if $f(x;\theta)$ can be written in the form:

$$f(x;\theta) = a(\theta)b(x)e^{c(\theta)d(x)} \text{ for } k_1 < x < k_2 \quad (1)$$

and zero elsewhere, for some functions $a(\theta)$, $b(x)$, $c(\theta)$, $d(x)$, where the constants k_1 and k_2 do not depend on θ; $a(\theta)$ and $c(\theta)$ are continuous and differentiable functions of θ; $c(\theta)d(x) = \sum_{i=1}^{t} c_i(\theta)d_i(x)$ and $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_s)$ for some t and s. Let $(X_1,\ldots,X_n)$ be a random sample of size n from $f(x;\theta)$ and let θ_1 and θ_2 be parametric values in Θ. We abbreviate $f(x;\theta_1) = f_1$ and $f(x;\theta_2) = f_2$. For two density functions $g_1(x)$ and $g_2(x)$, we define the *relative entropy* of $g_1(x)$ and $g_2(x)$ to be

$$D(g_1(x) \,\|\, g_2(x)) = \int_{-\infty}^{\infty} g_1(x) \ln \frac{g_1(x)}{g_2(x)} dx \quad (2)$$

where the support set of $g_1(x)$ is contained in the support set of $g_2(x)$. We consider testing $H_0 : \theta = \theta_1$ against $H_1 : \theta = \theta_2$ using the following acceptance region for H_0:

$$A_n(\alpha) = \left\{ (x_1,\ldots,x_n) : \frac{f(x_1,\ldots,x_n \mid \theta_1)}{f(x_1,\ldots,x_n \mid \theta_2)} > e^{\alpha n} \right\}, \quad (3)$$

where $-D(f_2 \,\|\, f_1) < \alpha < D(f_1 \,\|\, f_2)$ and $\dfrac{f(x_1,\ldots,x_n \mid \theta_1)}{f(x_1,\ldots,x_n \mid \theta_2)}$ is the likelihood ratio. Then the asymptotic behaviour of $\alpha_n = P(A_n^c(\alpha) \mid H_0)$ and $\beta_n = P(A_n(\alpha) \mid H_1)$ are given by:

$$\alpha_n \approx e^{-nD(f_\lambda \| f_1)} \quad (4)$$

and

$$\beta_n \approx e^{-nD(f_\lambda \| f_2)} \quad (5)$$

where

$$f_\lambda = f_\lambda(x;\theta_1,\theta_2) = r f^{\lambda}(x;\theta_1) f^{1-\lambda}(x;\theta_2) \quad (6)$$

and r is some normalizing constant for the density function f_λ (see [1]). The number λ satisfies $0 < \lambda < 1$ and the equation

$$D(f_\lambda \,\|\, f_2) - D(f_\lambda \,\|\, f_1) = \alpha . \quad (7)$$

For the case $\alpha = 0$, we have symmetry in the error exponents (4) and (5) and the common exponent $D(f_\lambda \,\|\, f_i)$ is called the *Chernoff information*. The aim of this paper is to determine the general forms of f_λ and $D(f_\lambda \,\|\, f_i)$ for the multi-parameter exponential class and present the Chernoff informations of some common distributions.

MAIN RESULTS

First, we have the following result.

Proposition 2.1. *Let $f(x;\theta)$ be a density function from the exponential class (1). Consider testing $H_0 : \theta = \theta_1$ against $H_1 : \theta = \theta_2$ using a likelihood ratio test with acceptance region for H_0 given by (3) and let $f_i = f(x;\theta_i)$ for $i = 1,2$. Then for large n, the two types of probabilities of error are given by:*

$$\alpha_n \approx e^{-nD(f_\lambda \| f_1)}$$

$$\textit{and} \qquad \beta_n \approx e^{-nD(f_\lambda \| f_2)}, \quad \textit{where}$$

$$f_\lambda = f_\lambda(x;\theta_1,\theta_2) = a_\lambda(\theta_1,\theta_2)b(x)e^{c_\lambda(\theta_1,\theta_2)d(x)} \quad \textit{for } k_1 < x < k_2, \quad (8)$$

$$c_\lambda(\theta_1,\theta_2) = c_\lambda = \lambda c(\theta_1) + (1-\lambda)c(\theta_2) \quad (9)$$

$$a_\lambda(\theta_1,\theta_2) = a_\lambda = \left[\int_{k_1}^{k_2} b(x)e^{c_\lambda(\theta_1,\theta_2)d(x)} dx \right]^{-1} \quad (10)$$

and $0 < \lambda < 1$ is a solution to the equation:

$$\ln\left[\frac{a(\theta_1)}{a(\theta_2)}\right] + E_{c_\lambda}\{[c(\theta_1) - c(\theta_2)]d(X)\} = \alpha , \quad (11)$$

where α is the exponent of the threshold function given in (3). The error exponent $D(f_\lambda \,\|\, f_i)$ is given by

$$D(f_\lambda \,\|\, f_i) = \ln\left[\frac{a_\lambda}{a(\theta_i)}\right] + E_{c_\lambda}\{[c_\lambda - c(\theta_i)]d(X)\} \quad \textit{for } i = 1,2 \quad (12)$$

where

$$E_{c_\lambda}[d(X)] = \left(\frac{-1}{\lambda a_\lambda}\right)\nabla a_\lambda(c(\boldsymbol{\theta}_1)) = \left[\frac{-1}{(1-\lambda)a_\lambda}\right]\nabla a_\lambda(c(\boldsymbol{\theta}_2)), \qquad (13)$$

$$\nabla a_\lambda(c(\boldsymbol{\theta}_1)) = \left[\frac{\partial a_\lambda}{\partial c_1(\boldsymbol{\theta}_1)},\ldots,\frac{\partial a_\lambda}{\partial c_t(\boldsymbol{\theta}_1)}\right] \quad \textit{and}$$

$$\nabla a_\lambda(c(\boldsymbol{\theta}_2)) = \left[\frac{\partial a_\lambda}{\partial c_1(\boldsymbol{\theta}_2)},\ldots,\frac{\partial a_\lambda}{\partial c_t(\boldsymbol{\theta}_2)}\right].$$

Proof. From (6),

$$f_\lambda(x;\theta_1,\theta_2) = rf^\lambda(x;\boldsymbol{\theta}_1)f^{1-\lambda}(x;\boldsymbol{\theta}_2)$$

for some normalizing constant r and hence

$$f_\lambda = f_\lambda(x;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = ra^\lambda(\boldsymbol{\theta}_1)a^{1-\lambda}(\boldsymbol{\theta}_2)b(x)e^{[\lambda c(\theta_1)+(1-\lambda)c(\theta_2)]d(x)}$$

for $k_1 < x < k_2$. Now, we can write

$$f_\lambda = a_\lambda(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)b(x)e^{c_\lambda(\theta_1,\theta_2)d(x)}$$

for $k_1 < x < k_2$, where the new normalizing constant

$$a_\lambda(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \left[\int_{k_1}^{k_2} b(x)e^{c_\lambda(\theta_1,\theta_2)d(x)}dx\right]^{-1}$$

and $c_\lambda(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \lambda c(\boldsymbol{\theta}_1) + (1-\lambda)c(\boldsymbol{\theta}_2)$. For simplicity, we shall suppress the vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $a_\lambda(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ and $c_\lambda(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ and write them as a_λ and c_λ, respectively. Now,

$$D(f_\lambda \,\|\, f_i) = \int_{k_1}^{k_2} a_\lambda b(x)e^{c_\lambda d(x)} \ln\left[\frac{a_\lambda b(x)e^{c_\lambda d(x)}}{a(\boldsymbol{\theta}_i)b(x)e^{c(\theta_i)d(x)}}\right]dx$$

$$= \ln\left[\frac{a_\lambda}{a(\boldsymbol{\theta}_i)}\right] + E_{c_\lambda}\{[c_\lambda - c(\boldsymbol{\theta}_i)]d(X)\} \quad \text{for } i = 1,2.$$

From (7), $0 < \lambda < 1$ is a solution to the equation:

$$D(f_\lambda \,\|\, f_2) - D(f_\lambda \,\|\, f_1) = \ln\left[\frac{a(\boldsymbol{\theta}_1)}{a(\boldsymbol{\theta}_2)}\right] + E_{c_\lambda}\{[c(\boldsymbol{\theta}_1) - c(\boldsymbol{\theta}_2)]d(X)\} = \alpha.$$

It remains to obtain an expression for $E_{c_\lambda}[d(X)]$ by differentiating the following integral with respect to $c(\boldsymbol{\theta}_1)$ componentwise:

$$\int_{k_1}^{k_2} a_\lambda b(x)e^{c_\lambda d(x)}dx = 1. \qquad (14)$$

As a result, we obtain

$$\frac{\partial a_\lambda}{\partial c_j(\boldsymbol{\theta}_1)}\frac{1}{a_\lambda} + \lambda E_{c_\lambda}[d_j(X)] = 0$$

for $j = 1,\ldots,t$, or

$$E_{c_\lambda}[d(X)] = \left(\frac{-1}{\lambda a_\lambda}\right)\nabla a_\lambda(c(\boldsymbol{\theta}_1)).$$

Similarly, differentiating (14) with respect to $c(\boldsymbol{\theta}_2)$ componentwise, we obtain

$$E_{c_\lambda}[d(X)] = \left[\frac{-1}{(1-\lambda)a_\lambda}\right]\nabla a_\lambda(c(\boldsymbol{\theta}_2)).$$

Remark. The normal distribution with p.d.f.

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = \frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{\mu^2}{2\sigma^2}}e^{\left(-\frac{1}{2\sigma^2}\right)x^2+\left(\frac{\mu}{\sigma^2}\right)x'}$$

belongs to the 2-parameter exponential class with $a(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{\mu^2}{2\sigma^2}}$, $b(x) = 1$,

$$c(\mu,\sigma) = \left[-\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2}\right] \text{ and } d(x) = [x^2, x].$$

The Chernoff informations $C(f_1,f_2)$ for some common distributions in testing on the given parameters are shown in Table 1.

Table 1. Chernoff informations of some common distributions.

Test and $D(f_1 \parallel f_2)$	f_λ	$C(f_1, f_2)$
$H_0 : f_1 = N(\mu_1, \sigma^2)$ $H_1 : f_2 = N(\mu_2, \sigma^2)$ $D(f_1 \parallel f_2) = \dfrac{(\mu_1 - \mu_2)^2}{2\sigma^2}$	$N(\lambda\mu_1 + (1-\lambda)\mu_2, \sigma^2)$	$D(f_\lambda \parallel f_1) = \dfrac{(\mu_2 - \mu_1)^2}{8\sigma^2}$
$H_0 : f_1 = N(\mu, \sigma_1^2)$ $H_1 : f_2 = N(\mu, \sigma_2^2)$ $D(f_1 \parallel f_2) = \ln\left(\dfrac{\sigma_2}{\sigma_1}\right) + \dfrac{[\sigma_1^2 - \sigma_2^2]}{2\sigma_2^2}$	$N(\mu, \tau^2)$ where $\tau^2 = \dfrac{\sigma_1^2 \sigma_2^2}{[\lambda(\sigma_2^2 - \sigma_1^2) + \sigma_1^2]}$	$D(f_\lambda \parallel f_1) = \ln\left(\dfrac{\sigma_1}{\tau}\right) + \dfrac{[\tau^2 - \sigma_1^2]}{2\sigma_1^2}$ where $\tau^2 = \left[\dfrac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2}\right] \ln\left(\dfrac{\sigma_1}{\sigma_2}\right)$
$H_0 : f_1 = \exp(\mu_1)$ $H_1 : f_2 = \exp(\mu_2)$ $D(f_1 \parallel f_2) = \ln\left(\dfrac{\mu_1}{\mu_2}\right) + \dfrac{(\mu_2 - \mu_1)}{\mu_1}$	$\exp[\mu_1\lambda + \mu_2(1-\lambda)]$	$D(f_\lambda \parallel f_1) = \ln\left(\dfrac{\mu_\lambda}{\mu_1}\right) + \dfrac{(\mu_1 - \mu_\lambda)}{\mu_\lambda}$ where $\mu_\lambda = \mu_1\lambda + \mu_2(1-\lambda)$, $\lambda = \dfrac{1}{\ln\left(\dfrac{\mu_1}{\mu_2}\right)} + \dfrac{\mu_2}{(\mu_2 - \mu_1)}$
$H_0 : f_1 = \text{Po}(\mu_1)$ $H_1 : f_2 = \text{Po}(\mu_2)$ $D(f_1 \parallel f_2) = (\mu_2 - \mu_1) + \mu_1 \ln\left(\dfrac{\mu_1}{\mu_2}\right)$	$\text{Po}(\mu_\lambda)$ where $\mu_\lambda = \mu_1^\lambda \mu_2^{1-\lambda}$	$D(f_\lambda \parallel f_1) = (\mu_1 - \mu_\lambda) + \mu_\lambda \ln\left(\dfrac{\mu_\lambda}{\mu_1}\right)$ where $\lambda = \dfrac{\ln\left[\dfrac{(\mu_1 - \mu_2)}{\mu_2 \ln\left(\dfrac{\mu_1}{\mu_2}\right)}\right]}{\ln\left[\dfrac{\mu_1}{\mu_2}\right]}$
$H_0 : f_1 = \text{geo}(p_1)$ $H_1 : f_2 = \text{geo}(p_2)$ $D(f_1 \parallel f_2) = \ln\left(\dfrac{p_1}{p_2}\right)$ $+ \left(\dfrac{1 - p_1}{p_1}\right) \ln\left(\dfrac{1 - p_1}{1 - p_2}\right)$	$\text{geo}(p_\lambda)$ where $p_\lambda = [(1 - p_1)^\lambda (1 - p_2)^{1}$	$D(f_\lambda \parallel f_1) = \ln\left(\dfrac{p_\lambda}{p_1}\right)$ $+ \left(\dfrac{1 - p_\lambda}{p_\lambda}\right) \ln\left(\dfrac{1 - p_\lambda}{1 - p_1}\right)$ where $p_\lambda = \dfrac{\ln\left[\dfrac{1 - p_1}{1 - p_2}\right]}{\ln\left[\dfrac{p_2(1 - p_1)}{p_1(1 - p_2)}\right]}$
$H_0 : f_1 = \text{bin}(m, p_1)$ $H_1 : f_2 = \text{bin}(\text{m}, p_2)$	$\text{bin}(m, p_\lambda)$ where $p_\lambda = \dfrac{\theta}{1 + \theta}$ and $\theta = \left[\dfrac{p_1}{1 - p_1}\right]^\lambda \left[\dfrac{p_2}{1 - p_2}\right]^{1-}$	$D(f_\lambda \parallel f_1) = m\left[p_\lambda \ln\left(\dfrac{p_\lambda}{p_1}\right)\right.$ $\left. + (1 - p_\lambda) \ln\left(\dfrac{1 - p_\lambda}{1 - p_1}\right)\right]$

$D(f_1 \parallel f_2) = m\left[p_1 \ln\left(\frac{p_1}{p_2}\right) + (1-p_1)\ln\left(\frac{1-p_1}{1-p_2}\right)\right]$		where $p_\lambda = \frac{\theta}{1+\theta}$ and $\theta = \frac{\ln\left[\frac{1-p_2}{1-p_1}\right]}{\ln\left[\frac{p_1}{p_2}\right]}$

ACKNOWLEDGEMENT

The author thanks the referee for his comments which motivated him to generalise the previous results to the multi-parameter exponential class.

REFERENCES

1. Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* New York: John Wiley.
2. Csiszar, I. and Shields, P. C. (2004). *Information Theory and Statistics: A Tutorial.* Hanover (Massachusetts): Now Publishers.