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 Bayes 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:
\( f(x; \theta) = a(\theta)b(x)e^{c(\theta)d(x)} \) for \( k_1 < x < k_2 \) \hspace{1cm} (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 \( \theta \); \( a(\theta) \) and \( c(\theta) \) are continuous and differentiable functions of \( \theta \);
\[
c(\theta)d(x) = \sum_{i=1}^{l} c_i(\theta)d_i(x) \quad \text{and} \quad \theta = (\theta_1, \ldots, \theta_s)
\]
for some \( t \) and \( s \). Let \( (X_1, \ldots, X_n) \) be a random sample of size \( n \) from \( f(x; \theta) \) and let \( \theta_1 \) and \( \theta_2 \) be parametric values in \( \Theta \). 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
\[
D(g_1(x) \parallel g_2(x)) = \int_{-\infty}^{\infty} g_1(x) \log \frac{g_1(x)}{g_2(x)} \, dx \hspace{1cm} (2)
\]
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 \( H_0 \):
\[
A_n(\alpha) = \left\{ (x_1, \ldots, x_n) : \frac{f(x_1, \ldots, x_n ; \theta_1)}{f(x_1, \ldots, x_n ; \theta_2)} > e^{n\alpha} \right\},
\]
where
\[
-D(f_2 \parallel f_1) < \alpha < D(f_1 \parallel f_2) \quad \text{and} \quad \frac{f(x_1, \ldots, x_n ; \theta_1)}{f(x_1, \ldots, x_n ; \theta_2)} \hspace{1cm} (3)
\]
the likelihood ratio. Then the asymptotic behaviour of \( \alpha_n = P(A_n(\alpha) \mid H_0) \) and \( \beta_n = P(A_n(\alpha) \mid H_1) \) are given by:
\[
\alpha_n \approx e^{-nD(f_2 \parallel f_1)} \hspace{1cm} (4)
\]
and
\[
\beta_n \approx e^{-nD(f_1 \parallel f_2)} \hspace{1cm} (5)
\]
where
\[
f_{\hat{\lambda}} = f_{\hat{\lambda}}(x; \theta_1, \theta_2) = r f_{\lambda}(x; \theta_1) f^{1-\lambda}(x; \theta_2) \hspace{1cm} (6)
\]
and \( r \) is some normalizing constant for the density function \( f_{\hat{\lambda}} \) (see [1]). The number \( \lambda \) satisfies \( 0 < \lambda < 1 \) and the equation
\[
D(f_{\hat{\lambda}} \parallel f_2) - D(f_{\hat{\lambda}} \parallel f_1) = \alpha \hspace{1cm} (7)
\]
For the case \( \alpha = 0 \), we have symmetry in the error exponents (4) and (5) and the common exponent \( D(f_{\hat{\lambda}} \parallel f_i) \) is called the Chernoff information. The aim of this paper is to determine the general forms of \( f_{\hat{\lambda}} \) and \( D(f_{\hat{\lambda}} \parallel f_i) \) 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:
\[
\alpha_n \approx e^{-nD(f_2 \parallel f_1)}
\]
and
\[
\beta_n \approx e^{-nD(f_1 \parallel f_2)}
\]
where
\[
f_{\hat{\lambda}} = f_{\hat{\lambda}}(x; \theta_1, \theta_2) = a_{\hat{\lambda}}(\theta_1, \theta_2) b(x) e^{c_{\hat{\lambda}}(\theta_1, \theta_2)d(x)} \hspace{1cm} (8)
\]
for \( k_1 < x < k_2 \),
\[
c_{\lambda}(\theta_1, \theta_2) = c_{\lambda}(\theta_1) + (1 - \lambda)c_{\theta_2} \hspace{1cm} (9)
\]
and \( 0 < \lambda < 1 \) is a solution to the equation:
\[
\ln \left[ \frac{a(\theta_1)}{a(\theta_2)} \right] + E \left[ (c(\theta_1) - c(\theta_2))d(X) \right] = \alpha \hspace{1cm} (10)
\]
where \( \alpha \) is the exponent of the threshold function given in (3). The error exponent \( D(f_{\hat{\lambda}} \parallel f_i) \) is given by
\[
D(f_{\hat{\lambda}} \parallel f_i) = \ln \left[ \frac{a_{\hat{\lambda}}}{a(\theta_i)} \right] + E \left[ (c_{\lambda} - c(\theta_i))d(X) \right] \hspace{1cm} (11)
\]
for \( i = 1, 2 \).
where

\[ E_{c_1}[d(X)] = \left( -\frac{1}{\lambda a_2} \right) \nabla a_2(c(\theta_2)) \]

\[ = \left( \frac{1}{1 - x_{\lambda}} \right) \nabla a_2(c(\theta_2)), \]  \hspace{1cm} (13)

\[ \nabla a_2(c(\theta_1)) = \frac{\partial a_2}{\partial c_1(\theta_1)} \nabla a_2(c(\theta_2)) \]

and

\[ \nabla a_2(c(\theta_2)) = \frac{\partial a_2}{\partial c_1(\theta_2)} \nabla a_2(c(\theta_2)). \]

Proof. From (6),

\[ f_\lambda(x; \theta_1, \theta_2) = r \lambda^2 1^{1/2} (x; \theta_1) f_1 \]

for some normalizing constant \( r \) and hence

\[ f_\lambda = f_\lambda(x; \theta_1, \theta_2) \]

\[ = r \lambda^2 1^{1/2} (x; \theta_2) b(x) e^{1/2(x; \theta_2) d(x)} \]

for \( k_1 < x < k_2 \). Now, we can write

\[ f_\lambda = a_\lambda(x; \theta_1, \theta_2) b(x) e^{1/2(x; \theta_2) d(x)} \]

for \( k_1 < x < k_2 \), where the new normalizing constant

\[ a_\lambda(\theta_1, \theta_2) = \left[ \int_{k_1}^{k_2} b(x) e^{1/2(x; \theta_2) d(x)} dx \right]^{-1} \]

and \( c_\lambda(\theta_1, \theta_2) = \lambda c(\theta_1) + (1 - \lambda) c(\theta_2) \). For simplicity, we shall suppress the vectors \( \theta_1 \) and \( \theta_2 \) in \( a_\lambda(\theta_1, \theta_2) \) and \( c_\lambda(\theta_1, \theta_2) \) and write them as \( a_\lambda \) and \( c_\lambda \), respectively. Now,

\[ D(f_\lambda || f_\lambda) = \int_{k_1}^{k_2} a_\lambda b(x) e^{1/2(x; \theta_2) d(x)} \ln \left( \frac{a_\lambda b(x) e^{1/2(x; \theta_2) d(x)}}{a(\theta_1, \theta_2) b(x) e^{1/2(x; \theta_2) d(x)}} \right) dx \]

\[ = \ln \left( \frac{\sigma_{\theta_1}}{\sigma_{\theta_2}} \right) + E_{c_1}[(c_{\theta_1} - c(\theta_2))d(x)] \] \hspace{1cm} for \( i = 1, 2 \).

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

\[ D(f_\lambda || f_\lambda) - D(f_\lambda || f_\lambda) \]

\[ = \ln \left( \frac{\sigma_{\theta_1}}{\sigma_{\theta_2}} \right) + E_{c_1}[(c_{\theta_2})d(x)] \]

\[ = \alpha. \]

It remains to obtain an expression for \( E_{c_1,d(X)} \) by differentiating the following integral with respect to \( c(\theta_1) \) componentwise:

\[ \int_{k_1}^{k_2} a_\lambda b(x) e^{c(\theta_1) d(x)} dx = 1. \]  \hspace{1cm} (14)

As a result, we obtain

\[ \frac{1}{\partial c(\theta_1) a_\lambda} + \lambda E_{c_1,d(X)}[d_j(X)] = 0 \]

for \( j = 1, \ldots, t \), or

\[ E_{c_1,d(X)} = \left( \frac{1}{\lambda \sigma_{\theta_1}} \right) \nabla a_\lambda(c(\theta_1)). \]

Similarly, differentiating (14) with respect to \( c(\theta_2) \) componentwise, we obtain

\[ E_{c_1,d(X)} = \left( \frac{1}{\lambda \sigma_{\theta_2}} \right) \nabla a_\lambda(c(\theta_2)). \]

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

\[ f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

\[ = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2\sigma^2} \left( -\frac{1}{\sigma^2} \right)^2 \left( -\frac{\sigma^2}{\sigma^2} \right)^2} \]

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

\[ c(\mu, \sigma) = \left( \frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2} \right) \] \hspace{1cm} 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>
<thead>
<tr>
<th>Test and $D(f_1 \parallel f_2)$</th>
<th>$f_\lambda$</th>
<th>$C(f_1, f_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0: f_1 = N(\mu_1, \sigma_1^2)$</td>
<td>$N(\lambda \mu_1 + (1-\lambda) \mu_2, \sigma^2)$</td>
<td>$D(f_\lambda \parallel f_1) = \frac{(\mu_2 - \mu_1)^2}{8\sigma^2}$</td>
</tr>
<tr>
<td>$H_1: f_2 = N(\mu_2, \sigma_2^2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D(f_1 \parallel f_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0: f_1 = N(\mu, \sigma_1^2)$</td>
<td>$N(\mu, \tau^2)$ where $\tau^2 = \frac{\sigma_1^2 \sigma_2^2}{\lambda(\sigma_2^2 - \sigma_1^2) + \sigma_1^2}$</td>
<td>$D(f_\lambda \parallel f_1) = \ln \left( \frac{\sigma_1}{\tau} \right) + \frac{(\tau^2 - \sigma_1^2)}{2\sigma_1^2}$</td>
</tr>
<tr>
<td>$H_1: f_2 = N(\mu, \sigma_2^2)$</td>
<td></td>
<td>where $\tau^2 = \left[ \frac{2\sigma_2^2 \sigma_1^2}{\sigma_1^2 - \sigma_2^2} \right] \ln \left( \frac{\sigma_1}{\sigma_2} \right)$</td>
</tr>
<tr>
<td>$D(f_1 \parallel f_2) = \ln \left( \frac{\sigma_1}{\sigma_2} \right) + \frac{[\sigma_1^2 - \sigma_2^2]}{2\sigma_2^2}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0: f_1 = \exp(\mu_1)$</td>
<td>$\exp[\mu_1 \lambda + \mu_2 (1-\lambda)]$</td>
<td>$D(f_\lambda \parallel f_1) = \ln \left( \frac{\mu_1}{\mu_2} \right) + \frac{(\mu_1 - \mu_2)}{\mu_2}$</td>
</tr>
<tr>
<td>$H_1: f_2 = \exp(\mu_2)$</td>
<td></td>
<td>where $\mu_\lambda = \mu_1 \lambda + \mu_2 (1-\lambda)$, $\lambda = \frac{1}{\ln \left( \frac{\mu_1}{\mu_2} \right) + \frac{\mu_2}{\mu_1}}$</td>
</tr>
<tr>
<td>$D(f_1 \parallel f_2) = \ln \left( \frac{\mu_1}{\mu_2} \right) + \frac{(\mu_2 - \mu_1)}{\mu_1}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0: f_1 = \text{Po}(\lambda_1)$</td>
<td>Po $\left( \mu_\lambda \right)$ where $\mu_\lambda = \mu_1 \lambda + \mu_2 (1-\lambda)$</td>
<td>$D(f_\lambda \parallel f_1) = (\lambda_1 - \lambda_2) + \mu_\lambda \ln \left( \frac{\mu_1}{\mu_2} \right)$</td>
</tr>
<tr>
<td>$H_1: f_2 = \text{Po}(\lambda_2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D(f_1 \parallel f_2) = (\mu_2 - \mu_1) + \mu_1 \ln \left( \frac{\lambda_1}{\lambda_2} \right)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0: f_1 = \text{geo}(p_1)$</td>
<td>geo($p_\lambda$) where $p_\lambda = [(1-p_1)^\lambda (1-p_2)]^{1/\lambda}$</td>
<td>$D(f_\lambda \parallel f_1) = \ln \left( \frac{p_1}{p_2} \right) + \frac{1-p_\lambda}{p_\lambda} \ln \left( \frac{1-p_\lambda}{1-p_1} \right)$</td>
</tr>
<tr>
<td>$H_1: f_2 = \text{geo}(p_2)$</td>
<td></td>
<td>where $p_\lambda = \frac{p_2 (1-p_1)}{p_1 (1-p_2)}$</td>
</tr>
<tr>
<td>$D(f_1 \parallel f_2) = \ln \left( \frac{p_1}{p_2} \right) + \frac{1-p_\lambda}{p_\lambda} \ln \left( \frac{1-p_\lambda}{1-p_1} \right)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0: f_1 = \text{bin}(m, p_1)$</td>
<td>bin($m, p_\lambda$) where $p_\lambda = \frac{\theta}{1+\theta}$ and $\theta = \left[ \frac{p_1}{1-p_1} \right]^{1/\lambda} \left[ \frac{p_2}{1-p_2} \right]^{1/\lambda}$</td>
<td>$D(f_\lambda \parallel f_1) = m \left[ \frac{p_1 \ln \left( \frac{p_1}{p_2} \right)}{p_1} \right] + (1-p_\lambda) \ln \left( \frac{1-p_\lambda}{1-p_1} \right)$</td>
</tr>
<tr>
<td>$H_1: f_2 = \text{bin}(m, p_2)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[D(f_1 \| f_2) = m \left[ p_1 \ln \left( \frac{p_1}{p_2} \right) + (1 - p_1) \ln \left( \frac{1 - p_1}{1 - p_2} \right) \right] \]

where \( p_3 = \frac{\theta}{1 + \theta} \) 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