As Medidas de Divergência e a Distância de Bhattacharyya em Seleção de Sinais

Thomas Kailath

***Resumo*— A minimização da probabilidade de erro para determinar sinais ótimos é por vezes difícil de se obter. Consequentemente, várias medidas de desempenho subótimas que são mais fáceis de se obter e manipular têm sido estudadas. Neste artigo parcialmente tutorial, nós comparamos as propriedades de uma medida geralmente usada, a divergência, com uma nova medida que chamamos de Distância de Bhattacharyya. Esta nova medida de distância é por vezes mais fácil de se obter que a divergência. Nos problemas em que trabalhamos, ela dá resultados que são pelo menos tão bons quanto, e às vezes melhores, que os dados pela divergência.**

***Palavras-Chave*— Teoria da Comunicação, Detecção, Sinais Digitais, Enfraquecimento, Filtros Combinados, Relação Sinal-Ruído.**

***Abstract*— Minimization of the error probability to determine optimum signals is often difficult to carry out. Consequently, several suboptimum performance measures that are easier than the error probability to evaluate and manipulate have been studied. In this partly tutorial paper, we compare the properties of an often used measure, the divergence, with a new measure that we have called the Bhattacharyya distance. This new distance measure is often easier to evaluate than the divergence. In the problems we have worked, it gives results that are at least as good as, and are often better, than those given by the divergence.**

***Keywords*— Communication Theory, Detection, Digital Signals, Fading, Matched Filters, Signal-to-Noise Ratio.**

I. INTRODUÇÃO

Em problemas de comunicação e radar, os sinais ótimos são aqueles que minimizam a probabilidade de erro. Entretando, em muitos casos a minimização da probabilidade de erro para determinar um conjunto de sinais ótimos é normalmente impossível. Isso pode ser porque uma expressão analítica explícita para o erro de probabilidade é muito difícil de encontrar e, mesmo se a expressão puder ser encontrada, ela pode ser muito complicada para minmização numérica ou analítica. Portanto, é útil procurar por critérios de seleção de sinal que podem ser mais fracos que a probabilidade de erro mas são mais fáceis de encontrar e manipular. Maximizar a taxa de defleção a relação sinal-ruído (SNR) são exemplos de critérios mais fracos - em geral, maximizar a SNR pode não levar à menor probabilidade de erro, mas a SNR ainda pode ser usada porque é mais fácil de analisar e provê algum direcionamento em muitos casos. Na busca por critérios aceitáveis, a noção de uma distância entre duas distribuições de probabilidade é razoavelmente útil. As duas distribuições serão distribuições das observações em uma hipótese binária testando a situação e quanto mais separadas forem estas distribuições, esperamos que menor será a probabilidade de confundir uma com a outra. Portanto, várias medidas de distância têm sido estudadas como substitutas mais simples para a probabilidade de erro.

O que nós gostaríamos para tais medidas de distância é uma propriedade como nos moldes de: se a distância entre as duas distribuições é maior para um conjunto de sinais α que para um conjunto β, então a probabilidade de erro para o conjunto α é sempre menor que para o conjunto β. Isto é bom demais para se ter esperança - nós só podemos encontrar relações mais fracas entre medidas de distância e a probabilidade de erro.

Em estatística, o uso destas medidas de distância tem um longo histórico começando com o trabalho de Pearson (como citado em Tildesley [1]). Duas medidas de distância amplamente usadas em estatística são a estatística-D^2 de Mahalanobis [2], [3] e a função discriminante linear introduzida por Fisher [4]. Estas medidas são discutidas em detalhe em muitos textos estatísticos, e.g., Anderson [5] ou Rao [6]. Com o advento da teoria de informação de Shannon em 1948, a divergência, uma medida fortemente relacionada com a medida de informação logarítmica de Shannon, ficou popular, apesar de ter sido proposta antes do trabalho de Shannon por Jeffreys [7]. Várias propriedades e aplicações da divergência são notadas em Kullback [8].

Na literatura, a divergência tem sido recentemmente aplicada em vários problemas com variado grau de sucesso (como visto em Marill e Green [9], Grettenberg [10], Hingorani [11], e outros). Neste artigo, propomos outro critério para seleção de sinal baseado em uma estatística introduzida primeiramente num contexto[2] estatístico para Bhattacharyya [12]. Nos vários problemas de seleção de sinal que consideramos até agora, a distância de Bhattacharyya deu resultados melhores que a divergência. Portanto, parece uma medida útil para se estudar.

Poderíamos apontar que a distância de Bhattacharyya é na verdade um caso especial de uma distância mais geral introduzida por Chernoff [15]. Esta medida, que será discutida brevemente na Seção IV, é em geral mais próxima da probabilidade de erro que a distância de Bhattacharyya. Por outro lado, esta medida não é tão fácil de se calcular quanto a distância de Bhattacharyya, que, por esta razão, terá preferência neste artigo.

Primeiramente, iremos revisar brevemente algumas das propriedades da divergência e da distância de Bhattacharyya e então iremos comparar as duas medidas de várias maneiras.

Iremos nos limitar largamente a problemas com duas hipóteses. As hipóteses serão denotadas por $h_i, i = 1, 2$; e nós determinamos $p_i(x)dx$ = a probabilidade de uma observação x quando a hipótese h_i é verdadeira. A observação x será tomada, por simplicidade, como um vetor de N compo-

[2]Outra aplicação estatística desta estatística foi feita por Kakutani [13], que nota uma aparição anterior em um problema não estatístico (Hellinger [14]). Portanto, os nomes Hellinger e Kakutani são frequentemente associados com a estatística de Bhattacharyya.

nentes. [Os conceitos podem ser estendidos para o caso de observações no tempo contínuo (veja Seção V).]

A taxa de similaridade:

$$L(x) = \frac{p_1(x)}{p_2(x)}$$

terá papel importante em nossas discussões. De fato, tanto a divergência quanto a distância de Bhattacharyya são funcionais convexas da taxa de similaridade.

II. A DIVERGÊNCIA E A DISTÂNCIA DE BHATTACHARYYA

A divergência foi introduzida inicialmente por Jeffreys [7], [16]. Ela é definida como a diferença nos valores médios da taxa de similaridade logarítmica sob duas hipóteses.

$$J = E_1[\ln L(x)] - E_2[\ln L(x)] \tag{1}$$

onde

$$E_i[\ln L(x)] = \int [\ln L(x)] p_i(x) dx, i = 1, 2. \tag{2}$$

Estas expectativas de $\ln L(x)$ normalmente são chamadaas números de Kullback-Leibler (Kullback e Leibler [17]) e escritas

$$I(1,2) = E_1[\ln L(x)], I(2,1) = -E_2[\ln L(x)]. \tag{3}$$

Em geral

$$I(1,2) \neq I(2,1). \tag{4}$$

A divergência é uma forma simetrizada dos números de K-L

$$J = I(1,2) + I(2,1). \tag{5}$$

Nós calcularemos J em vários exemplos na Seção III. A divergência satisfaz todos os postulados para uma métrica (distância) exceto o da desigualdade triangular. Que diz que nós temos $J(1,2) = J(2,1), J(1,2) \geq 0$ com igualdade somente quando $p_1 = p_2$, mas $J(1,2) + J(2,3)$ pode ser menor que $J(1,3)$. Esta última frase pode ser checada tomando p_1, p_2, p_3 como distribuições normais com média 0 e variâncias 1, 4 e 5.

Várias outras propriedades da divergência - sua aditividade para observações independentes, seu comportamento sob transformações, suas relações com capacidade de canal, etc. - e várias aplicações em classificação e teste de hipóteses são dadas em Kullback [8].

Para este artigo, estamos interessados na propriedade da divergência relacionada com a seleção de sinal. Se temos dois conjuntos de sinais - ou de forma mais geral, dois conjuntos de parâmetros para as densidades $p_i(x), i = 1, 2$ - digamos α e β, podemos *ordená-las* por meio da divergência. Ou seja, podemos dizer que o conjunto de sinais α é melhor que o conjunto β se $J(\alpha)$, a divergência entre p_1 e p_2 para parâmetros α é maior que $J(\beta)$. Em comunicações e radar, a frase "o conjunto de sinais α é melhor que conjunto β" geralmente significa que "a probabilidade de erro (ou de forma mais geral, o risco de Bayes) é menor para o conjunto α do que para o sinal β". Agora a probabilidade de erro depende da distribuição total (e consequentemente de todos os momentos) da taxa de similaridade enquanto a divergência depende somente dos primeiros momentos (média) da taxa de similaridade. Portanto, em geral, se $J(\alpha) > J(\beta)$ [i.e., α é melhor que β sob o critério da divergência] não será verdade que $P_e(\alpha) < P_e(\beta)$. Consequentemente, o teorema de Karlin e Bradt [18] é surpreendente.

Se $J(\alpha) > J(\beta)$, existe um conjunto de probabilidades a priori $\pi = (\pi_1, \pi_2)$ para as duas hipóteses, para os quais:

$$P_e(\alpha, \pi) < P_e(\beta, \pi)^3 \tag{6}$$

Este é um resultado interessante, mas seria muito mais se pudéssemos afirmar algo mais do que apenas a existência de um bom conjunto de probabilidades a priori. Infelizmente, isto é difícil de fazer. Mesmo assim, o resultado de Karlin e Bradt encorajou Grettenberg [10] a aplicar o critério da divergência em vários problemas de seleção de sinais. Vários o seguiram.

Antes de discutirmos estes problemas de seleção de sinais, vamos introduzir a distância de Bhattacharyya. Primeiramente, definimos o coeficiente de Bhattacharyya para duas densidades $p_i(x)$ por

$$\rho = \text{coeficiente de Bhattacharyya } = \int \sqrt{p_1(x) p_2(x) dx} \tag{7}$$

Claramente ρ está entre 0 e 1. Nós podemos associar várias medidas de distância com este coeficiente. O que usaremos será [4]

$$B = \text{a distância de Bhattacharyya } = -\ln \rho \tag{8}$$

Claramente $0 \leq B \leq \infty$. Podemos mostrar que B não precisa obedecer a desigualdade triangular (considerando populações normais com médias zero e variâncias iguais a 1, 4 e 5). (Mas $\sqrt{1-\rho}$ obedece à desigualdade triangular). Nós iremos calcular B para várias distribuições nas Seções III e IV.

A distância de Bhattacharyya tem várias propriedades interessantes, algumas das quais iremos discutir depois, mas o que nos levou a investigar sua aplicação na seleção de sinais foi o fato de que B tem a propriedade descoberta por Karlin e Bradt para a divergência.

Se, para dois ou mais conjuntos de parâmetros de sistema α e β, temos $B(\alpha) > B(\beta)$ ou de forma equivalente $\rho(\alpha) < \rho(\beta)$, então existe um conjunto

$$\pi = (\pi_1, \pi_2) \tag{9}$$

de probabilidades a priori para as quais $P_e(\alpha, \pi) < P_e(\beta, \pi)$.

Este resultado, juntamente com o anterior de Karlin e Bradt para a divergência, segue facilmente um teorema provado por Blackwell [19] e atribuído por ele a um trabalho não publicado por H. Bohnenblust, L. Shapley, and S. Sherman. Nós temos

[3] $P_e(\alpha, \pi)$ é a probabilidade de erro com o conjunto de parâmetros α e probabilidade a priori π

[4] Outra possibilidade é a quantidade $\sqrt{1-\rho}$

A. Teorema (Blackwell)

[5]

$P_e(\alpha, \pi)$ será menor ou igual a $P_e(\beta, \pi)$ para todo π se e somente se

$$E_\alpha[\phi(L_\alpha)|h^{(2)}] \leq E_\beta[\phi(L_\beta)|h^{(2)}] \tag{10}$$

para todas as funções côncavas contínuas $\phi(L)$. $E_\alpha[\phi(L_\alpha)|h^{(2)}]$ é a esperança de $\phi(L)$ a hipótese $h^{(2)}$.

A prova de (9) agora é simples. Notamos que $\sqrt{L}$ é uma função côncava de L e que

$$\begin{aligned}\rho(\alpha) &= \int \sqrt{p_1(x,\alpha)\sqrt{p_2(x,\alpha)dx}} \\ &= \int \sqrt{\frac{p_1(x,\alpha)}{p_2(x,\alpha)}} p_2(x,\alpha)dx = E_\alpha[\sqrt{L_\alpha}|h^{(2)}]\end{aligned} \tag{11}$$

Similarmente

$$\rho(\beta) = E_\beta[\sqrt{L_\beta}|h^{(2)}]$$

Mas então $\rho(\alpha) < \rho(\beta)$ sugere $E_\alpha[\sqrt{L_\alpha}|h^{(2)}]E_\beta[\sqrt{L_\beta}|h^{(2)}]$, o que contradiz [10]. Assim sendo, o teorema de Blackwell mostra que se $\rho(\alpha) < \rho(\beta)$, então $P_e(\alpha, \pi)$ não pode ser estritamente menor que $P_e(\beta, \pi)$ para todo π, o que é o resultado (9).

III. Algumas aplicações em seleção de sinais

Grettenberg [10] usou a propriedade de comparação (6) da divergência J como a racionalização para usar J como uma ferramenta de seleção de sinais. Nós achamos que a distância de Bhattacharyya B tem uma propriedade similar (9) e isto nos leva a comparar J e B como ferramenta para seleção de sinal. Nós iremos considerar três exemplos estidados originalmente pelo método da divergência - dois foram tratados por Grettenberg [10] e um por Reiffen e Sherman [20].

A. Processos Gaussianos com Funções de Valor Médio Desiguais

O problema de detecção é escolher entre duas hipóteses

$$h_1 : x(t) = m_1(t) + n(t), h_2 : x(t) = m_2(t) + n(t), t \in T$$

onde $m_1(t)$ e $m_2(t)$ são funções conhecidas no tempo, $n(t)$ é uma função amostra de um processo Gaussiano de média zero e função de covariância $R(t, s)$, e $x(t)$ é uma observação na base da qual poderemos decidir entre h_1 e h_2 qual é verdadeira. Por simplicidade, iremos assumir que o ruído é branco, com

$$R(t, s) = \frac{N_0}{2}\delta(t - s)$$

e que os sinais $m_i(t), i = 1, 2$ têm energia igual

$$\int_T {m_i}^2(t)dt = E, i = 1, 2 \tag{12}$$

e um coeficiente de correlação μ definido por

$$\int_T m_1(t)m_2(t)dt = E\mu. \tag{13}$$

Agora alguns cálculos irão mostrar que

$$B = \frac{1}{8}J = \frac{1}{8}\frac{2E}{N_0}(1 - \mu) \tag{14}$$

tais que tanto a distância de Bhattacharyya B quanto a divergência J são maximizadas ao se escolher os sinais $m_i(1), i = 1, 2$ que sejam antipodais, ou seja

$$m_1(t) = -m_2(t) \quad ou \quad \mu = -1 \tag{15}$$

É amplamente conhecido que esta escolha de sinais é de fato a que minimiza a probabilidade de erro para todas as probabilidades a priori, tal que neste problema tanto o critério de seleção de sinais da divergência quanto o da distância de Bhattacharyya produzem sinais que são de fato também ótimos com base na probabilidade de erro. Esta feliz coincidência infelizmente não é universal, como o próximo exemplo mostra.

B. Processos Gaussianos com Funções de Covariância Desiguais

O problema é o seguinte: um dos dois sinais $m_i(t), i = 1, 2$ cada um com energia E/L

$$\int_T {m_i}^2(t)dt = E/L, i = 1, 2 \tag{16}$$

é transmitido por L canais onde em cada um o sinal é perturbado multiplicativamente por α_k e aditivamente por $n_k(t)$, tal que os sinais recebidos são

$$h^{(i)} : x_k(t) = \alpha_k m_2(t) + n_k(t), k = 1, ..., L, i = 1, 2^6 \tag{17}$$

Vamos assumir que α_k são variáveis aleatórias Gaussianas independentes e identicamente distribuidas com

$$E[\alpha_k] = 0, E[\alpha_k\alpha_j] = \sigma^2\delta_{kj}, k, j = 1, ..., L \tag{18}$$

Assumiremos que os ruídos aditivos serão Gaussianos, independentes e identicamente distribuidos com

$$E[n_k(t)] = 0, E[n_k(t)n_j(s)] = [\frac{N_0}{2}\delta(t - s)]\delta_{kj} \tag{19}$$

e independentes das variáveis aleatórias $[\alpha_k]$

$$E[\alpha_k] = 0, E[\alpha_k n_k(t)] = 0, k, j = 1, ..., L \tag{20}$$

[5] Notamos que este teorema é atribuído a Karlin e Bradt por Grettenberg (compare com Teorema 1 de Grettenberg [10]), enquanto que o Teorema 2 de Grettenberg deveria de fato ter sido creditado a Karlin e Bradt.

[6] No problema de comunicação, é fisicamente mais razoável é considerar $\alpha_k, x_k(t), m_2(t) e n_k(t)$ envelopes comlexos dos sinais reais correspondentes (de banda estreita). Por comparação com os resultados iniciais de Pierce [21] e de Grettenberg [10], nós assumiremos que o são. Entretanto B e J podem ser calculados e otimizados quer os sinais sejam de banda estreita ou não.

O problema na seleção de sinais exposta acima é como segue: se nós tivermos apenas um canal, há uma alta probabilidade do sinal transmitido ser drasticamente enfraquecido pelo ruído multiplicativo assumindo um valor baixo. Para combater esta perca de sinal, que normalmente é chamada 'perda devido ao esvanecimento', uma solução clássica (1923 - 1927) é usar um número de canais para prover diversidade. Se nós usarmos um número de canais com ruídos multiplicativos independentes, a probabilidade de que o sinal vai ser simultaneamente reduzido em todos os links pode se tornar pequena. Seria, portanto, desejável usar a maior diversidade de canais possível. Contudo, outro fator deve ser levado em consideração. A energia total do sinal é normalmente restrita digamos, a um valor E e, assim sendo, quanto mais canais usarmos menor vai ser a energia de sinal disponível por canal. Se muitos canais são usados, a energia de sinal por canal pode ser tão pequena que a maior diversidade de ganho provida pela diversidade de canais não compensará a maior degradação resultante no desempenho de cada canal. Portanto deve haver um número ótimo de canais para usar dado uma energia total de sinal E.

Vejamos agora quais resultados obtemos com o uso da divergência e de Bhattacharyya. Primeiramente para a divergência, alguns cálculos mostram que, se (por simplicidade) nós assumimos adicionalmente que os sinais são ortogonais, ou seja

$$\int_0^T m_1(t)m_2(t)dt = 0 \tag{21}$$

então a divergência é

$$J = \frac{R^2}{R+L}, R = 2E\sigma^2/N_0 \tag{22}$$

E claramente para um E fixado, isso é maximizado estabelecendo $L = 1$, o que significa colocar toda a energia em um único canal. Entretanto, nossa discussão anterior indica que isto não é algo razoável de se fazer quando R é grande. E de fato, utilizando certas técnicas de delimitação e por estimação numérica, Pierce [21] encontrou o número ótimo de canais como uma função da SNR, $R = 2E\sigma^2/N_0$, para o caso de probabilidades a priori iguais. Pierce mostrou que para um baixo SNR $R << 1$ o uso de um único canal é ótimo, mas para um alto SNR $R >> 1$, o número ótimo de caanais é dado (aproximadamente) por uma fórmula notavelmente simples [7]

$$L_{opt} = R/3 \tag{23}$$

onde $[x]$ significa "o maior número inteiro menor ou igual a x". A conclusão que extraímos disto é que, tomando as palavras de Grettenberg ([10], p. 275), "a amplitude de estatísticas de fonte para as quais o código de máxima divergência tem a menor probabilidade de erro que um código diferente nem sempre inclui o caso equiprovável." Nós devemos repetir que o teorema de Karlin-Brandt (6) nos mostra que a solução de máxima divergência ($L = 1$) deve ser a melhor, não importa quão grande é a SNR, para um determinado conjunto de probabilidades a priori. Infelizmente, o teorema não nos dá nenhuma informação sobre esse conjunto de probabilidades bom e em muitos problemas pode ser que este conjunto seja altamente distorcido [π_1 = unidade] e, portanto, altamente improvável.

A distância de Bhattacharyya neste problema pode ser calculada como

$$\rho = e^{-B} = (1+\frac{R}{L})^{L}/(1+\frac{R}{2L})^{2L}, R = \frac{2E\sigma^2}{N_0}$$

A diferenciação nos mostra que B é máximo quando

$$\frac{x^2}{(x+1)(x+2)} = \ln\frac{x^2+4x+4}{4x+4}, x = R/L$$

e a solução computacional desta equação fornece

$$L_{opt} = R/3.07$$

o que é a mesma da solução 23 encontrada por Pierce [21] da expressão para P_e, que incidentalmente é

$$P_e = \frac{1}{(2+R/L)^L}\sum_{k=0}^{L-1}\left(\begin{array}{c} L-1+k \\ k \end{array}\right)\left(\frac{1+R/L}{2+R/L}\right)^k$$

Podemos demonstrar [veja discussão após (49)] que assintoticamente a distância de Bhattacharyya essencialmente determina a probabilidade de erro em um sentido que

$$\lim_{P_e\to 0} P_e = c.e^{-B}, c = uma \quad constante$$

Portanto, para um sinal Gaussiano (de média zero) no problema de ruído Gaussiano, a distância de Bhattacharyya fornece melhores resultados (pelo menos para a situação de maior interesse em problemas de comunicação onde $\pi_1 \doteq \pi_2$) que a divergência.

Processos Gaussianos com diferentes médias e covariância fornecem resultados similares - a divergência e a distância de Bhattacharyya fornecem resultados similares em uma baixa SNR[8], mas com uma alta SNR eles fornecem soluções diferentes, com a solução dada pela divergência sendo mais pobre (ao menos para $\pi_1 \doteq \pi_2$).

Contudo, como último exemplo, tratamos um problema em que os dois critérios fornecem resultados similares.

C. Processos de Poisson com Funções de Valor Médio Desiguais

O problema de detecção é escolher entre dois processos de Poisson cujas funções de valor médio são

$$m^{(k)}(t) = \lambda_i{}^{(k)} + \lambda_0, (i-1)\triangle \le t \le i\triangle, i = 1, ..., Nk = 1, 2 \tag{24}$$

[7] Grettenberg [10] erroneamente afirma que a probabilidade de erro para este caso diminui monotonicamente com L, tal que a melhor escolha é $L = \infty$

[8] É fácil checar neste problema que em uma baixa SNR, $J \doteq 8B$ [para qualquer valor de SNR, nós mostraremos (72) que para densidades Gaussianas $J \ge 8B$.

onde o intervalo de observação T foi dividido em N intervalos de comprimento $\triangle, N\triangle = T$. Os $\{{\lambda_i}^{(t)}\}$ são determinados pela nossa escolha de sinais. Nós restringimos essa escolha por

$$ {\lambda_i}^{(k)} \geq 0, \sum \triangle({\lambda_i}^{(1)} + {\lambda_i}^{(2)}) = 2E \tag{25} $$

A observação é a sequência $\{n_i\}$ do número de eventos (chegada de fóton em um detector de fótons) ocorridos em vários intervalos. o número de eventos no $i$ésimo intervalo de Poisson é distribuido com [9]

$$ \begin{aligned} &P^{(k)}\{n_i \text{ eventos em } (i-1)\triangle \leq t \leq i\triangle\} \\ &= \frac{[({\lambda_i}^{(k)} + \lambda_0)\triangle]^{n_i}}{n_i!} e^{-\triangle}({\lambda_i}^{(k)} + \lambda_0) \end{aligned} \tag{26} $$

Assumimos que os números de eventos em intervalos desjuntos são independentemente distribuidos. Para este problema, a divergência J e a distância de Bhattacharyya podem ser dadas por

$$ \begin{aligned} J = &\sum_1^N ({\lambda_i}^{(1)} + \lambda_0) \ln \left[\frac{({\lambda_i}^{(1)} + \lambda_0)}{({\lambda_i}^{(2)} + \lambda_0)} \right] \\ &+ \sum_1^N ({\lambda_i}^{(2)} + \lambda_0) \ln \left[\frac{({\lambda_i}^{(2)} + \lambda_0)}{({\lambda_i}^{(1)} + \lambda_0)} \right] \end{aligned} \tag{27} $$

$$ B = 2 \sum_1^N \left[\sqrt{{\lambda_i}^{(1)} + \lambda_0} - \sqrt{{\lambda_i}^{(2)} + \lambda_0} \right]^2 \tag{28} $$

Iremos primeiramente determinar a escolha de sinais (ou seja, dos ${\lambda_i}^{(k)}$) que maximizam J. O primeiro passo é claramente fazer

$$ {\lambda_i}^{(k)} = 0, quando \quad {\lambda_i}^{(j)} \neq 0, j \neq k \tag{29} $$

ou seja, fazer os sinais ortogonais em um certo sentido. Entretanto, a divergência pode ser ainda mais incrementada. Portanto, definamos a sequência

$$ q_i = {\lambda_i}^{(1)} se {\lambda_i}^{(2)} = 0, {\lambda_i}^{(1)} \geq 0 \tag{30} $$

A restrição (25) agora é $\sum_1^N q_i = 2E$ e a divergência é

$$ \tag{31} $$

$$ \frac{1}{\lambda_0} J = \sum \frac{q_i}{\lambda_0} \ln \left(\frac{q_i}{\lambda_0} + 1 \right), \sum q_i = 2E \tag{32} $$

Um ponto fixo para a expressão em (32) é facilmente encontrado usando os multiplicadores de Lagrange como

$$ q_i = 2E/N, i = 1, ..., N $$

Contudo, este ponto fixo é mais provavelmente um mínimo do que um máximo. Nós podemos mostrar que J é a função convexa de q_i tal que o máximo de J vai ocorrer em qualquer ponto extremo.

$$ q_i = 2E \quad para \quad alguns \quad i, q_i = 0, i \neq j \tag{33} $$

Portanto, uma escolha ótima de sinais (baseada no critério da divergência) é

$$ \{{\lambda_i}^{(1)}\} = \{2E, 0, ..., 0\}, \{{\lambda_i}^{(2)}\} = \{0, 0, ..., 0\} \tag{34} $$

ou qualquer permutação de (44).

Como mostraremos agora, a maximização de B fornece o mesmo resultado. Com a fórmula (28) para B em mente, notamos que

$$ \begin{aligned} &\left[\sqrt{{\lambda_i}^{(1)} + \lambda_0} - \sqrt{{\lambda_i}^{(2)} + \lambda_0} \right]^2 \\ &= ({\lambda_i}^{(1)} + \lambda_0) \left[1 - \sqrt{\frac{{\lambda_i}^{(1)} + \lambda_0}{{\lambda_i}^{(2)} + \lambda_0}} \right]^2 \\ &\leq ({\lambda_i}^{(1)} + \lambda_0) \left[1 - \sqrt{\frac{\lambda_0}{{\lambda_i}^{(1)} + \lambda_0}} \right]^2 \end{aligned} \tag{35} $$

com igualdade se e somente se ${\lambda_i}^{(2)} = 0$; O argumento pode ser repetido com ${\lambda_i}^{(1)}$ no lugar de ${\lambda_i}^{(2)}$. Portanto, para maximizar B, devemos estabelecer

$$ {\lambda_i}^{(k)} = 0, se {\lambda_i}^{(j)} > 0, j \neq k \tag{36} $$

Esta é também uma das condições necessárias para a maximização da divergência J. Portanto, como visto em (30), vamos definir uma sequência

$$ q_i = \begin{cases} {\lambda_i}^{(1)}, se {\lambda_i}^{(2)} = 0, {\lambda_i}^{(1)} \geq 0 \\ {\lambda_i}^{(2)}, se {\lambda_i}^{(1)} = 0, {\lambda_i}^{(2)} \geq 0 \end{cases} \tag{37} $$

com isso a distância de Bhattacharyya (28) pode ser escrita como

$$ B = 2 \sum_1^N [\sqrt{q_i + \lambda_0} - \sqrt{\lambda_0}]^2, \sum_1^N q_i = 2E \tag{38} $$

e assim temos

$$ B \leq 2[\sum_1^N \sqrt{q + \lambda_0} - \sqrt{\lambda_0}]^2 \tag{39} $$

com igualdade se e somente se todos os termos na soma (38) menos um, digamos $\sqrt{q + \lambda_0} - \sqrt{\lambda_0}$ são iguais a zero. Portanto uma escolha ótima de sinais (baseada no critério da distância de Bhattacharyya) é

$$ \{{\lambda_i}^1\} = \{2E, 0, ..., 0\} \{{\lambda_i}^2\} = \{0, 0, ..., 0\} \tag{40} $$

ou qualquer permutação desta.

Assim, os sinais considerados ótimos pelos critérios da divergência e da distância de Bhattacharyya são os mesmos. Alguns estudos computacionais por Blunlenthal (Stanford) indicam que sinais semelhantes a um pulso forte de forma

[9] Por conveniência, deste ponto em diante consideramos $\triangle = 1$

(34) ou (40) também minimizam a probabilidade de erro (para sinais igualmente prováveis). Nós devemos apontar que Reiffen e Sherman [20] sugeriram que a otimização de sinais como (34) ou (40) em baixa relação sinal ruído [10] essencialmente usando o critério da divergência. Abend [22] provou que sinais similares maximizaram uma relação sinal ruído previamente definida em todos os níveis de sinal.

IV. Algumas outras propriedades de B e J

Na última Seção, mostramos em vários problemas que a distância de Bhattacharyya é uma boa medida para seleção de sinais. Nesta seção, apresentaremos algumas propriedades adicionais de B, e traremos a comparação com a divergência J um pouco mais longe. Esta seção é de forma geral uma compilação de vários resultados selecionados de literatura em estatística. Iremos começar com a conveniente interpretação geométrica de B ou, na verdade, $\rho = e^{-B}$.

A. Interpretação Geométrica

Bhattacharyya [12] propôs que os números $\{\sqrt{p_i(x)}, \quad \text{todos} \quad \text{x} \quad \text{permissíveis}, i = 1, 2\}$ sejam considerados cossenos direcionais de dois vetores no espaço de x. Alternativamente, podemos considerá-los como dois pontos que definem a esfera unitária (como $\int p_i(x)dx = 1, i = 1, 2$) de x. O coeficiente de Bhattacharyya ρ pode ser considerado o cosseno do ângulo entre estes dois vetores, i. e., $\rho = cos\triangle, \triangle =$ ângulo entre as linhas com cossenos em diferentes direções.

$$\{\sqrt{p_i(x)}\} \tag{41}$$

O ângulo $\triangle$ deve claramente ser entre 0 e $\pi/2$, portanto $0 < \rho < 1$. Agora é natural considerar também a distãncia, digamos ∂, entre dois pontos $\{\sqrt{p_i(x)}, i = 1, 2\}$ na esfera unitária

$$\partial \triangleq \int \left[\sqrt{p_1(x)} - \sqrt{p_2(x)}\right]^2 dx \tag{42}$$

Claramente

$$\partial = 4 \sin \frac{\epsilon}{2} = 2(1 - \cos \triangle) = 2(1 - \rho) \tag{43}$$

A distância ∂ ocorre no trabalho de Jeffreys [7], [16].

B. A Distância Variacional de Kolmogorov

Outra medida de distância fortemente relacionada a ρ é a distância variacional de Kolmogorov (veja Adhikari e Joshi [23], p.71); ela é definida por

$$K(\pi) = \frac{1}{2} \int (|\pi_1 p_1(x) - \pi_2 p_2(x)|dx) \tag{44}$$

onde π e π_2 são as probabilidades a priori das hipóteses h_1 e h_2. Kraft [24] obteve certas desigualdades importantes entre ρ e $K_{(\pi)}$. Suas desigualdades (estendidas para o caso de $\pi_1 \neq \pi_2$) são

[10]Em baixa relação sinal ruído todas as medidas propostas parecem coincidir.

$$\sqrt{1 - 4\pi_1\pi_2\rho^2} \geq 2K(\pi) \geq 1 - \sqrt{\pi_1\pi_2\rho} \tag{45}$$

A primeira desigualdade vem da desigualdade de Schwartz

$$[\int |\pi_1 p_1 - \pi_2 p_2|dx]^2 \leq \int |\sqrt{\pi_1 p_1} - \sqrt{\pi_2 p_2}|^2 dx . \times$$
$$|\sqrt{\pi_1 p_1} - \sqrt{\pi_2 p_2}|^2 dx = [1 - 2\sqrt{\pi_1\pi_2\rho}][1 + 2\sqrt{\pi_1\pi_2\rho}]$$

A segunda desigualdade vem de

$$\int |\pi_1 p_1 - \pi_2 p_2|dx \geq \int |\sqrt{\pi_1 p_1} - \sqrt{\pi_2 p_2}|^2 dx$$
$$= 1 - 2\sqrt{\pi_1\pi_2\rho}$$

Há outro jeito útil de escrever $K(\pi)$. Um argumento padrão em análise mostra que

$$K(\pi) = \max_A |\pi_1 Pr\{x \in A|h_1\} - \pi_2 Pr\{x \in A|h_1\}| \tag{46}$$

onde o máximo é medido de todos os conjuntos A no espaço de x. Esta forma de $K(\pi)$ é fortemente relacionada com a probabilidade de erro (ótima). Para, se definirmos $A_i = \{x|h_i$ é escolhido quando x é recebido $\}$

$$\begin{aligned} P_e &= \min_{A_1}\{\pi_1 \int_{A_2} p_1(x)dx - \int_{A_1} p_2(x)dx\} \\ &= \pi_1 - \max_{A_1}\{|\int_{A_1} \pi_1 p_1(x) - \pi_2 p_2(x)dx|\} \\ &= \pi_1 - \max_{A_1} |Pr\{x \in A_1|h_1\} \equiv \pi_2 Pr\{x \in A_1|h_2\}| \\ &= \pi_1 - K(\pi) \end{aligned} \tag{47}$$

C. Limites Superiores e Inferiores em Probabilidade de Erro

Combinando (47) e as relações (45), notamos que

$$\begin{aligned} 2\pi_1 - \sqrt{1 - 4\pi_1\pi_2\rho^2} \leq 2P_e =& 2\pi_1 - K(\pi) \\ &\leq 2\pi_1 - 2\sqrt{\pi_1\pi_2\rho} \end{aligned} \tag{48}$$

se $\pi_1 = \pi_2 = \frac{1}{2}$, isso se torna

$$\frac{1}{4}\rho^2 \leq \frac{1}{2}(1 - \sqrt{1 - \rho}) \leq P_e \leq \frac{1}{2}\rho \text{[11]} \tag{49}$$

Os limites (49) não são inteiramente satisfatórios. À medida que $\rho \to 0$ (baixo P_e) a diferença entre os limites superior e inferior tende a zero mas sua razão tende ao infinito. Contudo, em um importante caso especial o limite superior é exponencialmente melhor, ou seja

$$\ln P_e \doteq -\ln \rho = B, P_e \to 0$$

Isso ocorre sob as seguintes circunstâncias (Chernoff [15]):

1) Os componentes X_i da observação $X' = X_1, ..., X_N$ são independentes e identicamente distribuídos.

[11]Se assumimos que $\pi_1 = \pi_2$ quando de fato eles não o são, nós só aumentamos a probabilidade de erro. Portanto, o limite superior em (49) é verdadeiro pra quaisquer π_1, π_2

2) $\int [p_1(x)]^{1-t}[p_2(x)]^t dx$ é simétrico em t.
3) N é grande.

Nós não iremos provar este resultado aqui. Devemos notar que resultados similares são encontrados no estudo de limites de probabilidade de erros para canais discretos e sem memória (Gallager [25]). O limite superior em (49) parece ser bom mesmo em algumas casos em que as condições que acabamos de dar não são encontradas (veja, por exemplo, o Exemplo B na última Seção), mas não parece ser possível fazer nenhuma afirmação geral.

D. A Divergência e a Probabilidade de Erro

O limite superior P_e em termos de ρ é bem útil. Nenhum limite similar em razão da divergência J parece ser geralmente verdadeiro. Contudo, usando uma desigualdade entre ρ e J podemos obter um limite inferior bruto de P_e em termos de J. As desigualdades básicas (que foram aparentemente derivadas primeiro por Hoeffding e Wolfowitz [26]) são

$$2B = -\ln \rho \leq I(1,2) \text{ e } 2B \leq I(2,1) \tag{50}$$

onde $I(1,2)$ e $I(2,1)$ são os números de Kullback-Leibler (K-L) definidos anteriormente [veja (2)-(3)]. A partir de (50), temos

$$4B \leq J \text{ ou } \rho \geq \exp(J/4) \tag{51}$$

As provas de (50) seguem pela igualdade de Jensen

$$-\frac{1}{2}I(1,2) = \int p_1(x) \ln \sqrt{\frac{p_2(x)}{p_1(x)}} dx = E_1[\ln \ \ sqrt\frac{p_2(x)}{p_1(x)}]$$
$$\ln[E_1(\sqrt{p_2(x)/p_1(x)})] = \ln \rho$$

Um limite inferior bruto de P_e agora parte de (49) e (51)

$$P_e \geq \rho^2/8 \geq \frac{1}{8}e^{J/2} \tag{52}$$

Claro que, em casos particulares, por exemplo quando $p_1(.)$ e $p_2(.)$ são Gaussianas, podemos obter relações mais estreitas entre B e J do que as providas por (51). Poderemos notar algumas destas relações especiais quando dermos fórmulas explícitas para B e J abaixo. Para fechar a presente discussão sobre limites de erro, notamos que a divergência e os números de K-L são úteis em prover limites inferiores nas probabilidades de erro condicionais. Estas seguem as duas fórmulas dadas por Kullback [8].

$$I(1,2) \geq P_e^1 \ln(\frac{P_e^1}{1-P_e^2}) + (1-P_e^1)\ln(\frac{1-P_e^1}{P_e^2}) \tag{53}$$

e

$$I(2,1) \geq P_e^2 \ln(\frac{P_e^2}{1-P_e^1}) + (1-P_e^2)\ln(\frac{1-P_e^2}{P_e^1}) \tag{54}$$

Note que para um valor fixo de $P_{e1}\{P_{e2}\}$, (53), (54) porvê um limite inferior em $P_{e1}\{P_{e2}\}$. O RHS em (53) e (54) têm interpretações interessantes - eles são os valores de $I(1,2)$ e $I(2,1)$ para distribuições binomiais com parâmetros P_{e1} e $1-P_{e2}$. Tabelas para cálculos em (53) e (54) são dadaw por Kullback [8], (veja pag. $74-75$ e $378-379$).

Ao passo que com a distância de Bhattacharyya, os limites providos por (53) e (54) são exponencialmente ótimos sob certas condições. Portanto se $x' = x_1, ..., x_N$ onde os $x_i, i = 1, ..., N$ são independentes e identicamente distribuídos, foi mostrado que

$$\text{para qualquer valor de } P_e^{(1)}, \lim_{N\to\infty} -\frac{\ln P_e^{(2)}}{N} = I(1,2)$$

onde

$$I(1,2) = \int p_1(x) \ln[\frac{p_1(x)}{p_2(x)}]dx$$

Uma afirmação similar se mantém mesmo com 1 e 2 trocados. Estes resultados foram dados primeiramente por Stein (não publicado); uma prova aparece em Kullback [8]. (Veja também Chernoff [27]). Daly [28] aplicou estes resultados de *design* de sinal de radar.

E. Relação de ρ com Informação de Fisher

Em problemas de estimação de parâmetros, nós normalmente temos não apenas duas, mas um contínuo de hipóteses. Para tais problemas, uma medida de informação especial, introduzida por Fisher em 1925 tem sido usada na literatura estatística, veja por exemplo Kullback [8], p. 26. Acontece que a distância de Bhattacharyya e a divergência são reduzidas a problemas de estimação de informação de Fisher. Para mostrar isso, tomemos $p(x|\theta) =$ uma função densidade de probabilidade dependendo de um parâmetro nos números reais $\theta, 0 \leq \theta \leq 1$, digamos. Voltando à nossa figura geométrica (41), nós podemos calcular

$$\begin{aligned}\cos \delta s &= \int \sqrt{p(x|\theta)p(x|\theta+\delta\theta)}dx \\ &\doteq 1 - \frac{(\delta\theta)^2}{8}\int_{-\infty}^{\infty}\frac{[\dot{p}(x|\theta)]^2}{p(x|\theta)}dx\end{aligned} \tag{55}$$

em que $\dot{p}(x|\theta) = (d/dx)\dot{p}(x|\theta)$ e $\delta s =$ é a distância real entre pontos com cossenos de direção

$$\sqrt{p(x|\theta)} \text{ e } \sqrt{p(x|\theta+\delta\theta)} \tag{56}$$

No limite enquanto δs e $\delta\theta$ tendem a zero nós obtemos

$$\begin{aligned}(\frac{ds}{d\theta})^2 &= \frac{1}{4}\int_{-\infty}^{\infty}\frac{[\dot{p}(x|\theta)]^2}{p(x|\theta)}dx \\ &= -\frac{1}{4}\int_{-\infty}^{\infty} p(x|\theta)[\frac{\partial^2}{\partial\theta^2}\ln p(x|\theta)]d\theta\end{aligned} \tag{57}$$

Esta derivação é devida a Bhattacharyya [12]. Um cálculo similar para a divergência[12] pode ser encontrado em Kullback [8], pag. 55. A aparição de certo modo inesperada da informação de Fisher em certos problemas de detecção singular foi notada recentemente por Shepp [29].

F. Fórmulas Explícitas para B e J

Nós podemos calcular B explicitamente para uma grande classe de distribuição - a família exponencial. Antes de dar a fórmula geral nós listaremos alguns casos especiais importante.

1) Distribuições Multinomiais:

$$\begin{aligned} &\text{Se } p_1(x) = \sum_{j=1}^{N} p_j^{(i)} \delta(x-j), \\ &\text{então } B = -\ln[\sum_{j=1}^{N} \sqrt{p_j^{(1)} p_j^{(2)}}] \end{aligned} \tag{58}$$

2) Distribuições de Poisson:

$$\begin{aligned} &\text{Se } p_1(x) = e^{m_i \cdot \frac{(m_i)^n}{n!}} \delta(x-n), \\ &\text{então } B = -\frac{1}{2}(\sqrt{m_1} - \sqrt{m_2})^2 \end{aligned} \tag{59}$$

3) Distribuições Gaussianas Univariadas:

$$\begin{aligned} &\text{Se } p_1(x) = N(\mu_1, \sigma_1), \\ &\text{então } B = \frac{1}{4}\frac{(m_1-m_2)^2}{\sigma_1^2 - \sigma_2^2} + \frac{1}{2}\ln\{\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1\sigma_2}\} \end{aligned} \tag{60}$$

4) Distribuições Gaussianas Multivariadas:

$$\begin{aligned} &\text{Se } p_1(x) = N(\mu_1, R_1), \\ &\text{então } B = \frac{1}{8}(m_1-m_2)' R^1 (m_1-m_2) \\ &\qquad + \frac{1}{2}\ln\{\frac{detR}{\sqrt{detR_1.detR_2}}\} \end{aligned} \tag{61}$$

em que

$$2R = R_1 + R_2 \tag{62}$$

Huzurbazar [30] calculou B para densidades exponenciais.

5) Densidades Exponenciais: Digamos que $p_i(x)$

$$= d(x)b(h_i)\exp[\sum_{j-1}^{r} V_j(h_i)U_j(x)], i = 1, 2. \tag{63}$$

As $V_i(h_i)$ são funções independentes de h_i, ou mais precisamente, de V parâmetros de densidade $p_i(.)$. Como $b(h_i)$ também depende de h_i, nós podemos expressar $b(h_i)$ em termos de $V_i(h_i)$, como

$$b(h_i) = g\{[V_j(h_i)]\} \tag{64}$$

[Nota: g depende de h_i apenas através de $V_j(h_i), j = 1..., r$]. Então a distância de Bhattacharyya é

$$B = -\ln\rho, \rho = \frac{\sqrt{b(h_j)b(h_i)}}{g(\{\frac{1}{2}V_j(h_1) + \frac{1}{2}V_j(h_2)\})} \tag{65}$$

[12]A variedade da medida de distância parece ter uma propriedade similar para pequenas variações

6) A Divergência: Huzurbazar [30] também calcula a divergência J para as densidades exponenciais

$$J = \sum_{j=1}^{r}[V_j(h_2) - V_j(h_1)][E_2V_j(x) - E_1V_j(x)] \tag{66}$$

onde

$$\begin{aligned} J = -\sum_{j=1}^{r}\sum_{k=1}^{r} \frac{\partial^2}{\partial V_j \partial V_k} \times \\ \{V_j(h_2) - V_j(h_1)\}\{V_k(h_2) - V_k(h_1)\} \end{aligned} \tag{67}$$

Nós damos um caso especial por causa de sua importânca. Se $p_i(x) = N(\mu_1, R_1)$, então

$$\begin{aligned} I(1,2) = &\frac{1}{2}\ln\frac{detR_2}{detR_1} + \frac{1}{2}trR_1[R_2^{-1} - R_1^{-1}] + \\ &\frac{1}{2}trR_2^{-1}[m_1-m_2][m_1-m_2]' \end{aligned} \tag{68}$$

e

$$\begin{aligned} J(1,2) = &\frac{1}{2}tr[R_1-R_2][R_2^{-1} - R_1^{-1}] + \\ &\frac{1}{2}tr[R_1^{-1} + R_2^{-1}][m_1-m_2][m_1-m_2]' \end{aligned} \tag{69}$$

Se as covariâncias são iguais a, digamos, $R_1 = R_2 = R$, temos

$$2I(1,2) = J = trR^{-1}[m_1-m_2][m_1-m_2]'^{13} \tag{70}$$

Notamos que a distância de Bhattacharyya neste caso é

$$B = \frac{J}{8} \tag{71}$$

Quer ou não que $R_1 = R_2$, podemos estabelecer a desigualdade [que é melhor por um fator de dois que a desigualdade geral (51)]

$$J \geq 8B \tag{72}$$

A prova mais simples de (72) é obtida considerando o caso de $R_1 = \Lambda$, uma matriz diagonal com as entradas $X_i, i = 1, ..., N$ e $R_2 = I$, a matriz identidade. (O caso geral pode ser obtido a partir daí utilizando uma matriz não singular para diagonalizar simultaneamente R_1 e R_2.) Daí, a partir de (67) e (75) nós obtemos, após alguma álgebra,

$$8B - J = 2\sum \ln(1+\lambda)^2/4\lambda_i - \sum(\lambda_i - 1)^2/2\lambda_i$$

Mas da desigualdade $ln(1+\lambda) < \lambda$ nós temos

$$\ln(1+\lambda)^2/4\lambda = \ln(1 + [(\lambda-1)^2/4\lambda]) \leq (\lambda-1)^2/4\lambda \tag{73}$$

tal que, utilizando isto para cada X_i nós obtemos

$$8B - J \leq 0$$

[13]Esta é a famosa Estatística D^2 de Mahalanobis.

7) Equações Diferenciais para B e J: Recentemente tem aparecido interesse (Schweppe [31], Schweppe and Athans [32]), em aplicar técnicas de teoria de controle a design de sinais. Para fazer isto, é conveniente obter equações diferenciais (no tempo-contínuo) como critério de desempenho. Scweppe usou B e J para estes critérios e obteve equações diferenciais para eles quando o sinal e o ruído são projeções de processos de Gauss-Markov. As fórmulas detalhadas são muito complicadas para reproduzir aqui, mas podem ser encontradas em Schweppe [31]. Contudo, nós podemos notar que para processos Gaussianos com covariâncias desiguais as equações diferenciais para J são tão mais complicadas que as equações para B que Schweppe nem coloca equações para J. Para citá-lo, "leitores que compartilham da teoria que 'a melhor resposta é a mais simples' irão decidir que a distância de Bhattacharyya é superior à divergência."

V. Outras Aplicações e Extensões para Tempo-contínuo

A distância de Bhattacharyya encontrou várias aplicações em estatística clássica (veja, por exemplo, os artigos de Matusita [33], Hannan [34], Rao [35]. Uma aplicação recente é por Stein [36], pag. 17. O resultado de Stein é que o aumento no risco de Bayes em uma problema de decisão quando uma distribuição prior incorreta, digamos π', é usada em uma hipótese é limitado (sob condições de regularidade suficientes) por uma constante vezes a distância de Bhattacharyya entre π' e a distribuição correta π. Este resultado também se estende a distribuições priores com parâmetros desconhecidos.)

Nós tratamos apenas do caso de duas hipóteses aqui. No caso de M hipóteses, parece que a distância média de Bhattacharyya

$$\overline{B} = \sum_{i,j}^{M} \pi^{(i)}\pi^{(j)} \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx$$

seria uma medida aceitável para seleção de sinal. Esta medida média terá a mesma propriedade minimax discutida por Grettenberg [10] para a divergência média. Além disso parece (como visto em Gallager [25]) que , para baixos P_e, B essencialmente determina a probabilidade de erro, P_e, no caso de M hipóteses.

Nós fechamos com alguns comentários sobre o caso de tempo contínuo. O método usual de resolver problemas no tempo contínuo é começar com um problema aproximado no tempo discreto e obter a resposta final como o limite da solução do problema no tempo discreto. O mesmo procedimento funciona aqui com uma modificação. Se nós temos $X_{(t_i)} = x_i, i = 1, ..., N$, então a distância de Bhattacharyya quando as observações são restritas a $\{t_i, i = 1, ..., N\}$ é

$$B_N = -\ln \int \sqrt{p_{1N}(x)p_{2N}(x)}dx \tag{74}$$

onde

$$\begin{aligned} &p_{1N}(x) - \text{função densidade de probabilidade} \\ &\qquad \text{de } N \text{variáveis aleatórias} x(t_i), i = 1, ..., N \end{aligned} \tag{75}$$

Quando $N \to \infty, \lim_{N\to\infty} p_{iN}(x)$ normalmente não é definido.

Contudo, as razões

$$\lim_{N\to\infty} q_{iN}(x) = \lim_{N\to\infty} \frac{p_{iN}(x)}{p_{iN}(x) + p_{2N}(x)} \tag{76}$$

serão bem comportadas (se os conjuntos $\{t_i\}$ são monotonicamente crescentes), ao passo que $N \to \infty$. Contudo, no tempo contínuo devemos definir B como

$$B = -\ln \rho, \text{onde } \rho = \lim_{N\to\infty} \int \sqrt{q_{1N}(x)q_{2N}(x)}dx \tag{77}$$

Argumentos similares podem ser usados para definir J no caso de tempo contínuo. Com esta modificação, todas as relações e propriedades que obtivemos nas Seções III e IV permanecem também para o caso de observações no tempo contínuo. As relações

$$\frac{1}{8}\rho^2 \leq P_e \leq \frac{1}{2}\rho \tag{78}$$

provêm um critério simples para detecção singular, ou seja, permitem discriminar entre dois processos com uma probabilidade de erro arbitrariamente baixa. Nós vemos que uma condição suficiente, primeiramente notada por Kraft [24], para detecção singular (extrema) é a de $\rho = 0$. A distância de Bhattacharyya também pode ser usada para prover uma prova simples de um resultado de Hajek [37] que o problema de detecção para dois processos Gaussianos para os quais a divergência $J = \infty$ é singular. A prova parte essencialmente do fato que no caso Gaussiano nós podemos suplementar o limite inferior (51) em $J(J \geq 4B)$ por um limite superior em J na forma de $J \leq (constante) \times B$. Este e outros aspectos da interação entre B e J em questões de detecção singular serão discutidas em uma nota separada.

Referências

[1] Tildesley, "A first study of the Burmese skull", *Biometrika*, vol 13, pp. 176-262, 1921.

[2] P. C. Mahalanobis, "Analysis of race mixture in Bengal", *J. Asiat. Soc. (India)*, vol. 23, pp. 301-310, 1925.

[3] P. C. Mahalanobis, "On the generalized distance in statistics", *Proc. Natl. Inst. Sci. (India)*, vol. 12, pp. 49-55, 1936.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems", *Ann. Eugmcis*, vol. 7, pp. 179-188, 1936; *Contributions to Mathematical Statistics*. New York: Wiley, 1950.

[5] T. W. Anderson, *An Introduction to Multiuariate Statistical Analysis*. New York: Wiley, 1955.

[6] C. R. Rao, *Advanced Statisticul Methods in the Estimation of Statistical Parameters*. New York: Wiley, 1952.

[7] H. Jeffreys, "An invariant form for the prior probability in estimation problems", *Proc. Roy. Soc. A.*, vol. 186, pp. 453-461, 1946.

[8] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[9] T. Marill and D. M. Green, "On the effectiveness of receptors In recognition systems", *IEEE Trans. on Information Theory*, vol. IT-9, pp. 11-17, January 1963.

[10] T. L. Grettenberg, "Signal selection in communication and radar systems", *IEEE Trans. on Information Theory*, vol. IT-9, pp. 265-275, October 1963.

[11] G. P. Hingorani and J. C. Hancock, "A transmitted reference system for communication in random or unknown channels", *IEEE Trans. on Communication Technology*, vol. COM-13, pp. 293-301, September 1965.

[12] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by thew probability distributions", *Bull. Calcutta Math. Soc.*, vol, 35, pp. 99-109, 1943.
[13] S. Kakutani, "On equivalence of rnfinlte product measures", *Ann. Math. Stat.*, vol. 49, pp. 214-2?24, 1948.
[14] E. Hellinger, "Neue begrundung der theorie quadrátischer formen von uendlichvielen veränderlichen", *J. fur die Reine und angew. Math.*, vol. 36. pp. 210-271, 1909.
[15] H. Chernoff, "A measure of asymptotic efficiency for tests of hypothesis based on a sum of observations", *Ann. Math. Stat.*, vol. 23, pp. 493-507, 1952.
[16] H. Jeffreys, *Theory of Probability*. Oxford University Press, 1948.
[17] S. Kullback and 12. A. Leibler, "Ch inform:ition and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79-86, 1951.
[18] S. Karlin and R. N. Bradt, "On the design and comparison of dichotomous experlments," *Ann. Math. Stat.*, Val. 27, PP.
[19] D. Blackwell, "Comparison of experiments", *Proc. Second Berkeley Symp. on Probability and Statistics*. Berkeley, Calif.:University of California Press, vol. 1, pp. 93-102, 1951.
[20] B. Reiffen and H. Sherman, "An optimum demodulator for Poisson process: photon source detectors" *Proc. IEEE*, vol. 51, pp. 1316-1320, October 1963.
[21] J. N. Pierce, "Theoretical limitations of frequency and time diversity for fading binary transmissions", *IRE Trans. on Communications Systems (Corres.)*, vol. CS-9, pp. 186-157, June 1961. June 1961.
[22] K. Abend, "Optimum photon detection", *IEEE Trans. on Information Theory*, vol. IT-12, pp. 64-65, January 1966.
[23] [23] B. P. Adhikari and D. D. Joshi, "Distance discrimation et resume exhaustif", *Publs. Inst. Statist.*, vol. 5, pp. 57-74, 1956.
[24] C. H. Kraft, "Some conditions for consistency and uniform consistency of statistical procedures", *University of California Publications in Statistics*, 1955.
[25] R. G. Gallager, "A simple derivation of the coding theorem and some applications", *IEEE Trans. on Information Theory*, vol. IT-11, pp. 3-18, January 1965.
[26] W. Hoeffding and J. Wolfowitz, "Distinguishability of sets of distributions", *Ann. Math. Stat.*, vol. 29, pp. 700-718, 1958.
[27] H. Chernoff, "Large-sample theory: parametric case", *Ann. Math. Stat.*, vol. 27, pp. 1-22, 1956.
[28] R. F. Daly, "Signal design for efficient detection in randomly dispersive media", SRI Tech. Rept. on Project 186531-144, 1965.
[29] L. A. Shepp, "Distinguishing a sequence of random variables from a translate of itself", *Ann. Math. Stat.*, vol. 36, pp. 1107-1112, 1965.
[30] V. S. Huzurbazar, "Exact forms of some invariants for distributions admitting sufficient statistics", *Biometrika*, vol. 42, pp. 533-537, 1955.
[31] F. Schweppe, "On the distance between Gaussian processes: the state space approach", *Information and Control*, 1967.
[32] F. Schweppe and M. Athans, "On the design of optimal modulation schemes via control-theoretic concepts I: formulation", to be published.
[33] K. Matusita, "Decision rules, based on the distance for problems of fit, two samples and estimation", *Ann. Math. Stat.*,vol. 26, pp. 631-640, 1955.
[34] J . Hannan, "Consistency of maximum likelihood estimation of discrete distributions", in *Contributions to Probabiltty and Statistics*, I. Olkin, Ed. Stanford, Calif.: Stanford University Press, 1960.
[35] C. R. Rao, "Asymptotic efficiency and limiting information", *Proc. Fourth Berkeley Symp. on Probability and Statistics*. Berkeley, Calif.: University of California Press, vol. 1, pp. 531-545, 1961.
[36] C. Stein, "Approximation of improper prior measures bg Prior probability measures", Department of Statistics, Stanford University, Stanford, Calif., Tech. Rept. 12, 1964.
[37] J. Hajek, "On a property of normal distribution of any stochastic process" (in Russian), *Czech. Math. J.*, vel. 83, pp. 610-618, 1958; a translation appears in "Selected Translations" in *Math. Statistics und Probability*, vol. 1, pp. 245-252.