Universidade Estadual de Campinas

Instituto de Física "Gleb Wataghin"

Estudo Sobre a Aplicação de Estatística Bayesiana e Método de Máxima Entropia em Análise de Dados

Eder Arnedo Perassa

Este exemplar corresponde à redação final da tese de mestrado defendida pelo aluno Eder Arnedo Perassa e aprovada pela Comissão Julgadora

3/03/2008

J. S. Chinellato

Orientador:

Prof. Dr. José Augusto Chinellato

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IFGW - UNICAMP

P411e

Perassa, Eder Arnedo
Estudo sobre a aplicação de estatística bayesiana e método de máxima entropia em análise de dados / Eder Arnedo Perassa. -- Campinas, SP : 2007.

Orientador: José Augusto Chinellato.
Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Física "Gleb Wataghin".

1. Raios cosmicos. 2. Chuveiros de raios cosmicos. 3. Estatistica bayesiana. 4. Metodo de entropia maxima. I. Chinellato, José Augusto. II. Universidade Estadual de Campinas. Instituto de Física "Gleb Wataghin". III. Título.

(smcc/ifgw)

- **Título em inglês:** Study on application of bayesian statistics and method of maximum entropy in data analysis.
- **Palavras-chave em inglês (Keywords):**
  1. Cosmic rays
  2. Cosmic rays showers
  3. Bayesian statistics
  4. Maximun entropy method
- **Área de concentração:** Física das partículas elementares e campos
- **Titulação:** Mestre em física
- **Banca examinadora:**
  Prof. José Augusto Chinellato
  Profª. Márcia Begalli
  Prof. Jun Takahashi
- **Data da defesa:** 19.04.2007

MEMBROS DA COMISSÃO JULGADORA DA TESE DE MESTRADO DE **EDER ARNEDO PERASSA – RA 040820** APRESENTADA E APROVADA AO INSTITUTO DE FÍSICA "GLEB WATAGHIN", DA UNIVERSIDADE ESTADUAL DE CAMPINAS, EM 19 / 04 / 2007.

COMISSÃO JULGADORA:

Prof. Dr. José Augusto Chinellato (Orientador do Candidato) – DRCC/IFGW/UNICAMP

Profa. Dra. Márcia Begalli – IF/UERJ

Prof. Dr. Jun Takahashi – DRCC/IFGW/UNICAMP

Dedicado aos meus pais, Sonia e Luiz, e ao meu avô Américo (In Memoriam)

Agradecimentos

Antes de tudo, um agradecimento especial a minha familia, meu alicerce e porto seguro sem o qual eu nada seria: ao meu pai Luiz e a minha mãe Sonia; e também a minha irmã Lígia, minha avó, tios e primos.

Ao Prof. Dr. José Augusto Chinellato, por sua orientação nessa dissertação, suas idéias, sua paciência e auxílio nas dificuldades.

Aos meus grandes amigos de Andradina, Marcelo e Wagner, pelo companheirismo de longos anos que não se desfez mesmo quando nossas vidas foram tomando rumos que nos obrigou a andarem separados.

Aos amigos de Bauru, Raphael, Fabiano, Cleber, Tiago, Airton, Roberto, Douglas e Renata, pelos momentos divididos, tanto os de estudo quanto os de descontração.

Ao grande amigo Pedro, pela amizade e irmandade ao longo desses anos.

Aos componentes da república "Chinatown"em Campinas, Bráulio, Júnior e Neemias, pelas conversas, risadas e todos os bons momentos passados.

Aos professores da Unesp em Bauru: Prof. Dr. Renato Ghiotto, Prof. Dr. José Brás Barreto de Oliveira, Prof. Dr. Américo Tabata, Prof. Dr. Edson Sardella, pelo incentivo na iniciação à pesquisa em Física.

Aos membros das bancas do Exame de Qualificação, Pré-Requisito e da Defesa de Tese, pelos conselhos e sugestões para um melhor andamento do projeto.

A FAPESP e a CNPq pelo financiamento e bolsa de estudos.

"Os pequenos atos que se executam são melhores que todos aqueles grandes que apenas se planejam. "

George C. Marshall

Resumo

Neste trabalho são estudados os métodos de estatística bayesiana e máxima entropia na análise de dados. É feita uma revisão dos conceitos básicos e procedimentos que podem ser usados para inferência de distribuições de probabilidade. Os métodos são aplicados em algumas áreas de interesse, com especial atenção para os casos em que há pouca informação sobre o conjunto de dados. São apresentados algoritmos para a aplicação de tais métodos, bem como alguns exemplos detalhados em que espera-se servirem de auxílio aos interessados em aplicações em casos mais comuns de análise de dados.

Abstract

In this work, we study the methods of Bayesian Statistics and Maximum Entropy in data analysis. We present a review of basic concepts and procedures that can be used for inference of probability distributions. The methods are applied in some interesting fields, with special attention to the cases where there's few information on set of data, which can be found in physics experiments such as high energies physics, astrophysics, among others. Algorithms are presented for the implementation of such methods, as well as some detailed examples where it is expected to help interested in applications in most common cases of data analysis.

Sumário

Lista de Figuras

Lista de Tabelas

1

Introdução

Os métodos bayesianos e de máxima entropia vêm sendo utilizados atualmente em Física. Eles são úteis especialmente em situações nas quais as várias repetições de determinado experimento com o objetivo de diminuir a incerteza na medida são muito caros e consomem muito tempo, o que é bastante comum em experiências na astronomia e astrofísica [1] e na física de altas energias [2]. Além dessas, estão sendo empregados os métodos bayesianos em áreas como espectroscopia de massa [3], espalhamento Rutherford [4] e ressonância magnética nuclear [5] .

Pode-se dizer que o objetivo geral do processo de inferência indutiva é o de, a partir de uma distribuição de probabilidades a priori, atualizá-la em uma distribuição posterior quando novas informações tornam-se disponíveis. Dois métodos destinados a esse fim e que serão discutidos na presente dissertação são: o método baseado no teorema de Bayes e o outro baseado no princípio de maximização da entropia (MaxEnt), sendo a escolha entre os dois definida pela natureza da informação que está sendo processada.

O teorema de Bayes é utilizado em situações que envolvem a inferência sobre determinado parâmetro θ com base no conhecimento de valores observados de outras quantidades y - os dados - e em alguma relação conhecida entre y e θ. Por outro lado, quando a informação a ser incorporada ao problema toma forma de um vínculo em uma família de distribuições posteriores admissíveis utiliza-se o método de Máxima Entropia [6]

A definição bayesiana de probabilidade difere da ortodoxa, ou frequentista. Segundo a definição ortodoxa, probabilidade é definida como sendo a *frequência relativa de ocorrência de um evento*,

que se aplica a experimentos repetidos ou um *ensemble* de sistemas "identicamente preparados". A definição bayesiana é mais abrangente pois considera a probabilidade como a medida de plausibilidade de determinada proposição quando o nosso conhecimento incompleto não permite estabelecer a falsidade ou veracidade da proposição com certeza.

Não é objetivo dessa dissertação uma discussão sobre qual das definições é a "correta", mas tão somente apresentar o método de Bayes e o de Máxima Entropia e mostrar como ele pode ser utilizado como um método de inferência na física.

Para tanto, na sequência desse capítulo, fazemos um breve histórico de como as idéias em teoria de probabilidade foram evoluindo com o tempo, desde os primeiros trabalhos de Bernoulli.

No segundo capítulo, discutimos a estatística bayesiana, apresentamos algumas propriedades e aplicações em inferência de parâmetros de modelos físicos, mostrando alguns exemplos de modelos comuns em física, tais como o modelo de Poisson. Nesta parte apresentamos a nossa contribuição a essa área estudando o caso da estimativa de parâmetros de uma distribuição genérica. Ela foi especialmente escolhida por ocorrer como solução geral de vários tipos de equação de transporte, tanto de difusão de raios cósmicos no espaço quanto na atmosfera terrestre.

O terceiro capítulo dedicamos ao estudo do princípio de maximização da entropia. Iniciamos com a dedução da função da entropia de Shannon e as propriedades que a definem; em seguida apresentamos o método de Jaynes e sua ampliação para um conceito geral chamado entropia relativa. Mostramos que os métodos de Bayes e o do Máxima Entropia são consistentes e, por fim, apresentamos uma série de aplicações de obtenção de distribuições a partir do princípio de máxima entropia.

Por fim, apresentamos nossas conclusões e referências usadas na elaboração da presente dissertação.

1.1 Breve Histórico

Pode-se dizer que mesmo em nosso cotidiano, nos deparamos com a necessidade do uso dos métodos de inferência em situações onde nossa informação é incompleta. Em um nível intuitivo, antes de tomarmos qualquer decisão nossa intuição organiza nosso raciocínio preliminar em vários estágios: (I) Tentar prever todas as possibilidades que poderão surgir; (II) Julgar quão provável é cada uma baseado em todo o conhecimento que temos sobre o assunto; (III) Analisar quais as

possíveis consequências que as várias ações podem ter; (IV) em seguida tomar a decisão.

Desde tempos muito antigos o processo de raciocínio plausível para tomar decisões era conhecido. Heródoto, em aproximadamente 500 A.C. discutia as decisões políticas dos reis persas. Ele notou que uma decisão era sábia, mesmo quando levava a consequências desastrosas, se a evidência em mãos indicava que ela fosse a melhor a se tomar; e que uma decisão era tola, mesmo pensando que levasse a possíveis consequências melhores, se não havia motivo para esperar por tais consequências.

Provavelmente, o primeiro trabalho a estabelecer uma representação matemática para o problema foi feito por Jaymes Bernoulli(1713) [7], chamado "Ars Conjectandi". Neste artigo, Bernoulli criou uma maneira de expressar nosso estado de conhecimento incompleto enumerando um conjunto de casos "igualmente possíveis"que podemos denotar por $x_1, x_2, ..., x_N$ que podem ser eventos ou mesmo proposições, contidos em um espaço de hipóteses H_0, de dimensão N. Dada qualquer proposição de interesse A, definida como sendo verdadeira em algum subconjunto $H(A)$ de M pontos de H_0, a probabilidade de A é definida como sendo a proporção:

$$p(A) = \frac{M}{N} \tag{1.1}$$

Bernoulli também estabeleceu a primeira conexão matemática entre probabilidade e frequência (Lei fraca para grandes números). Se fizermos n observações independentes e encontrarmos A verdadeiro m vezes, a frequência $f(A) = m/n$ é comparada à probabilidade $p(A) = M/N$ e no limite de n tendendo a infinito são iguais. Thomas Bayes foi um clérigo britânico e matemático amador, que em 1763 publica um trabalho [8] chamado "Estatística Bayesiana". Ao contrário de Bernoulli que calculava a probabilidade que poderíamos observar A verdadeiro m vezes, dados os valores de N, n e M, Bayes propôs uma expressão para o cálculo para a probabilidade, a partir de N, n, m, de M resultar em algum determinado valor. O método ficou conhecido como "probabilidade inversa".

Em um de seus trabalhos publicados, Laplace [9] redescobriu o princípio de Bayes de uma forma mais clara e geral e aplicou-o a inúmeros problemas de astronomia, meteorologia, estatítica de populações e mesmo jurisprudência. Denotando as várias proposições por A,B,C, etc., e sendo AB a proposição de que ambos A e B sejam verdadeiros, $\bar{A}$ representando que A seja falso e que $P(A|B)$ significa como "a probabilidade que A seja verdadeiro dado que B é verdadeiro", as regras básicas

do produto e soma da teoria de probabilidade seriam escritas como:

$$p\left(AB|C\right) = p\left(A|BC\right)p\left(B|C\right) \tag{1.2}$$

$$p\left(A|B\right) + p\left(\bar{A}|B\right) = 1 \tag{1.3}$$

No entanto, AB e BA representam as mesmas proposições, portanto, podemos trocar A e B de lugar na equação (1.2) de modo a obtermos o que é conhecido hoje como sendo o Teorema de Bayes, mesmo Bayes nunca o tendo escrito dessa forma:

$$p\left(A,B|I\right) = p\left(A|I\right)\frac{p\left(B|A,I\right)}{p\left(B|I\right)} \tag{1.4}$$

O termo $p\left(A|I\right)$ é chamado de probabilidade a priori ("prior probability") de A, quando conhecemos unicamente a informação I, que foi escrita explicitamente no final de cada probabilidade para deixar claro que a inferência baseia-se em I que representa toda informação relevante que temos em mãos sobre A antes de obter B. O termo $p\left(A|B,I\right)$ é a probabilidade posterior, atualizada como resultado de termos obtido nova informação B. O termo que relaciona as duas, $p\left(B|A,I\right)$, é chamado de função de verossimilhança ("likelihood function"). Tipicamente, A representa uma hipótese que queremos testar, B representa novos dados a partir de observações.

Em um famoso exemplo de aplicação por Laplace do Teorema de Bayes a proposição A representava que o valor da massa desconhecida de Saturno M_S estava em um determinado intervalo específico, B os dados a partir das observações nas perturbações mútuas de Júpiter e Saturno, I a noção de que a massa de Saturno não poderia ser tão pequena a ponto de saturno perder seus anéis e não tão grande de modo que escapasse do Sistema Solar. Laplace reportou que usando o Teorema de Bayes estimou que M_S era de 1/3512 da massa solar, e deu uma probabilidade de .99991 que M_S possuía esse valor. Após um acúmulo de dados de mais de 150 anos a estimativa aumentou 0.63 porcento, o que é um resultado notável.

Muitos anos passaram-se até que o trabalho de Laplace fosse redescoberto e expandido por Sir Harold Jeffreys [10] nos anos 30 e em seguida por Richard T.Cox [11] que deduziu pela primeira vez as regras da probabilidade utilizando para isso condições de consistência em forma de equações funcionais, mostrando que as soluções de tais equações determinavam unicamente as regras obtidas

anteriormente por Laplace e Jeffreys. Assim, Cox mostrou por um teorema que qualquer método de inferência que representasse graus de plausibilidade por números reais seria necessariamente equivalente as regras de Laplace-Jeffreys, ou inconsistentes. Dois anos depois, Claude Shannon(1948) [12] usou o método de Cox novamente. Ele buscava uma medida da "quantidade de incerteza"em uma distribuição de probabilidade. Novamente, as condições de consistência tomaram a forma de equações funcionais cuja solução geral ele encontrou. A medida encontrada por ele foi:

$$S_I = -\sum p_i ln p_i \tag{1.5}$$

Gibbs(1875) chegou a um princípio variacional cuja maximização da "entropia de Clausius" S_E leva a todas as predições úteis da termodinâmica de não-equilíbrio. Mas S_E ainda tinha de ser determinada por medidas calorimétricas, o familiar cálculo $S_E = \int dQ/T$ sobre um caminho reversível. Boltzmann(1877), Gibbs(1902) e von Neumann(1928) chegaram a três princípios variacionais em cuja maximização da "entropia de informação de Shannon"levava, tanto na teoria clássica como na quântica, a predição teórica da entropia de Clausius e todos os resultados úteis da termodinâmica de equilíbrio, sem qualquer necessidade de medidas calorimétricas.

Em 1957, Jaynes [13] propôs o chamado "Princípio de Maximização da Entropia", pelo qual busca-se estimar uma distribuição de probabilidades $(p_1, p_2, ..., p_n)$ em algum espaço de hipótese usando por critério a maximização da entropia de informação de Shannon sujeita aos vínculos que expressam propriedades que desejamos que a distribuição possua, mas não são suficientes para determiná-la univocamente.

Após o desenvolvimento por Jaynes do método, houve a necessidade de ampliar a definição de entropia, primeiramente porque a definição de Shannon relacionava-se à quantidades discretas, nada dizendo sobre como tratar o caso contínuo e, segundo, pois buscava-se uma noção de entropia que estivesse fora de qualquer necessidade de justificação via definições como "quantidade de informação ou incerteza". Neste sentido, foram bastante importantes os trabalhos de Shore e Johnson [14] que ampliaram a definição de entropia para casos contínuos, obtendo a chamada entropia relativa.

2

Métodos Bayesianos

Vimos que o teorema de Bayes origina-se da simetria da regra do produto (1.2). O rearranjamento dessa expressão leva a:

$$P(H_j|E_i, I) = \frac{P(E_i|H_j, I)P(H_j|I)}{P(E_i|I)} \tag{2.1}$$

Aqui, H representa uma hipótese qualquer e E são os eventos ou dados. Obtivemos assim uma regra lógica para atualizar nossas crenças sobre determinada proposição com base em novas informações disponíveis. Na maioria dos casos práticos, o cálculo de $P(E_i|I)$ pode ser difícil, enquanto a determinação da probabilidade $P(E_i|H_j, I)$ é mais fácil. Podemos reescrever $P(E_i|I)$ da equação anterior em termos de quantidades presentes no numerador do seguinte modo:

$$P(E_i|I) = \sum_j P(E_i|H_j, I)P(H_j, I) \tag{2.2}$$

Deste modo, precebe-se que $P(E_i|I)$ é um fator de normalização e, assim, o teorema de Bayes é escrito como:

$$P(H_j|E_j, I) \propto P(E_i|H_j, I)P(H_j|I) \tag{2.3}$$

Se temos um conjunto de observações $\{d_i\}$, para os quais desejamos aplicar o teorema de Bayes afim de obter uma inferência sobre o valor de um parâmetro θ, podemos pensar em inferir θ de

acordo com a análise de cada novo dado d_i separadamente . Ou seja, aplicando o teorema de Bayes uma vez:

$$P(\theta|d_1, I) \propto P(d_1|\theta, I)P(\theta|I) \tag{2.4}$$

E depois da segunda observação, incorporamos um novo dado d_2 e novamente aplicamos o teorema de Bayes:

$$P(\theta|d_1, d_2, I) \propto P(d_2|\theta, d_1, I)P(\theta|d_1, I) \propto P(d_2|\theta, d_1, I)P(d_1|\theta, I)P(\theta|I) \tag{2.5}$$

Por outro lado, podemos aproveitar a regra do produto para escrever:

$$P(d_1, d_2|\theta, I) = P(d_2|\theta, d_1, I)P(d_1|\theta, I) \tag{2.6}$$

E assim,

$$P(\theta|d_1, d_2, I) \propto P(d_1, d_2|\theta, I)P(\theta|I) \tag{2.7}$$

Comparando com a (2.5) vemos que uma sequência de inferências leva ao mesmo resultado de uma única inferência que leva em conta toda a informação disponível de uma única vez, o que é um importante resultado do formalismo bayesiano [15].

A extensão para muitas variáveis resulta em:

$$P(\{\theta_i\} \mid \{d_i\}, I) \propto P(\{d_i\} \mid \{\theta_i\}, I)P(\{\theta_i\} |I) \tag{2.8}$$

Além disso, quando os d_i são independentes obtemos para a verossimilhança:

$$P(\{d_i\} \mid \{\theta_i\}, I) = \prod_i P(d_i| \{\theta_i\}, I) \tag{2.9}$$

ou seja, a verossimilhança combinada é dada pelo produto das verossimilhanças individuais.

2.1 Inferência de Valores Numéricos de Quantidades Físicas Usando o Método Bayesiano

Em física, estamos interessados em modelos ("teorias") e nas quantidades físicas relacionadas ambas as quais, em geral, são hipóteses que queremos inferir, dadas as observações. Por outro lado, em muitas aplicações há fortes razões para acreditar em qual modelo (representado no teorema de Bayes como a distribuição de verossimilhança) deve ser usado para interpretar as medidas. A seguir, apresentamos uma série de aplicações em estimativa de parâmetros, situações em que a escolha do modelo é ditada pela natureza do problema.

2.1.1 Exemplo da Moeda

Em análise de dados, muitas vezes estamos interessados em conhecer o valor de determinado parâmetro dada uma distribuição qualquer. Iniciaremos esta seção com um exemplo simples que ilustra bem como isso é feito, utilizando-se o método bayesiano. Nosso problema consiste em analisar um conjunto de lançamentos de uma moeda e verificar, por meio disso, se esta é uma moeda "honesta".

Por moeda honesta queremos dizer que dada um conjunto muito grande de lançamentos poderíamos esperar que os resultados de cara e coroa sejam em 50 : 50. Uma maneira de formular tal problema é considerar um conjunto de proposições ou hipóteses que representaria o "bias-weighting"ou "grau de honestidade"H da moeda, ou seja, se H=0 significa que a moeda produz somente lançamentos que dão coroa, H=1 apenas lançamentos que resultam em cara e $H = 0,5$ representaria uma moeda honesta.

Assim, há um contínuo de possibilidades para os valores de H. por exemplo, as proposições poderiam ser: $(a)0.00 \leq H < 0.01; (b)0.01 \leq H < 0.02$, e assim sucessivamente. Nosso grau de conhecimento, então, sobre a honestidade da moeda seria completamente especificado conhecendo o quanto acreditamos que cada proposição seja verdadeira. Se ao final da análise, estiver atribuída uma alta probabilidade à determinada proposição (ou um pequeno conjunto delas), então podemos estar confiantes que temos uma boa estimativa do parâmetro H; caso contrário, estando as várias proposições atribuídas com probabilidades equivalentes, então não temos condições de determinar com muita precisão o valor de H e, portanto, se a moeda é honesta ou não. De acordo com o

Teorema de Bayes, temos:

$$p\left(H|\left\{dados\right\},I\right) \propto p\left(\left\{dados\right\}|H,I\right) \times p\left(H|I\right) \tag{2.10}$$

A probabilidade a priori, $p\left(H|I\right)$, representa o nosso conhecimento sobre a hipótese, no caso, a "honestidade"da moeda, antes que qualquer dado seja analisado. Vamos adotar uma total ignorância sobre a veracidade da hipótese H, ou seja, admitiremos que a probabilidade de obter qualquer valor de H seja a mesma, independemente do seu valor:

$$p\left(H|I\right) = cte. \tag{2.11}$$

sendo que $0 \leq H \leq 1$.

Esta probabilidade é modificada pelos dados através da função de verossimilhança, $p\left(\left\{dados\right\}|H,I\right)$. Ela mede a probabilidade de obtermos os valores que observamos, dado um valor de H conhecido. Se, em nossa informação a priori I, assumimos que os lançamentos da moeda são eventos indepententes, então o resultado de um lance não interfere no próximo valor.

Em situações como essa, usa-se o chamado *modelo binomial* cuja aplicabilidade se extende em uma grande classe de processos físicos em que as observações consistem de contagens, ou seja, número de eventos ou ocorrências. A distribuição binomial descreve a probabilidade de obter n eventos (sucessos)em um total de N eventos estatisticamente independentes, em que assumimos a mesma probabilidade que cada evento ocorra, o que no nosso caso representaria a probabilidade de obter R caras em N lances. Assim:

$$p\left(\left\{dados\right\}|H,I\right) \propto H^{R}(1-H)^{N-R} \tag{2.12}$$

Deste modo, a probabilidade posterior vai sendo modificada à medida que os dados são computados. Para tornar o exemplo mais claro, reportamo-nos aos resultados da simulação computacional de dados, apresentadas em [16].

Supondo que a moeda seja lançada e o resultado seja "cara", a probabilidade posterior é modificada para uma reta inclinada (Fig. 2.1).

$$p\left(H|\left\{dados\right\},I\right) \propto H \tag{2.13}$$

Se o segundo resultado da simulação for novamente "cara", iremos obter uma probabilidade posterior proporcional a H^2, como pode ser visto no gráfico(ver Fig. 2.1, terceiro painel). No terceiro lançamento, obtém-se "coroa"pela primeira vez e, portanto, a probabilidade de que $H = 1$ é nula. No quarto lançamento novamente obtém-se "coroa"; assim, calculando o ponto de máximo da probabilidade posterior, $p\left(H|\left\{dados\right\},I\right) = H^2(1-H)^2$, encontramos $H = 0.5$, poderíamos chegar à conclusão de que a moeda seja "honesta". No entanto, há uma grande incerteza envolvida nessa estimativa, como podemos perceber observando a largura da distribuição em torno do ponto de máximo (ver Fig. 2.1, segundo painel inferior).

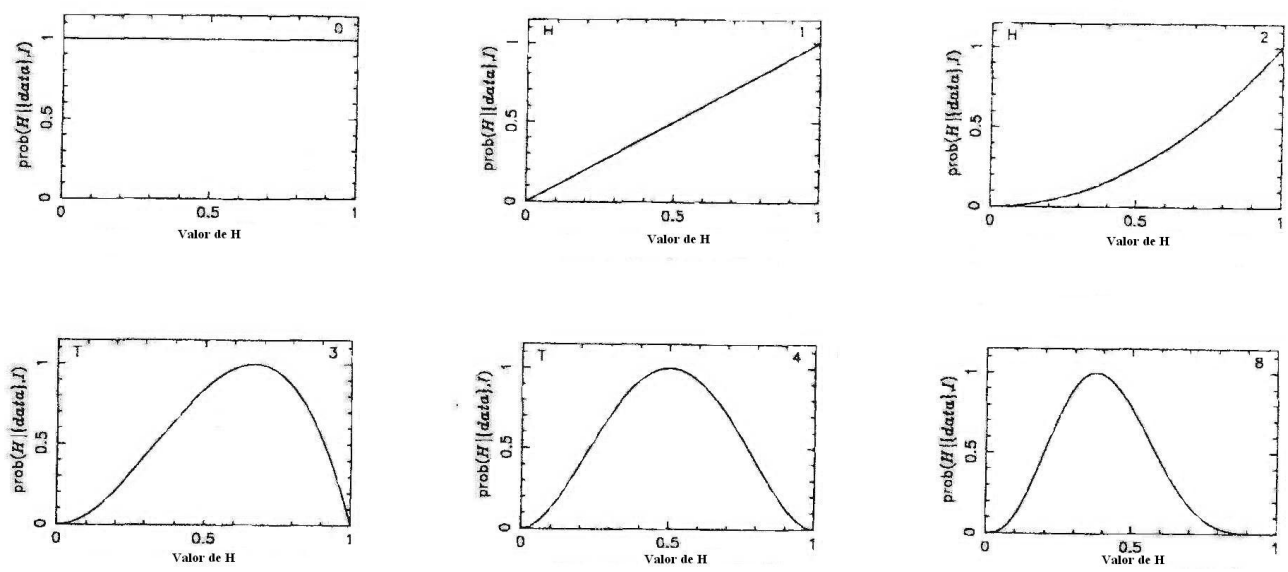

Figura 2.1: Evolução da distribuição posterior do valor de H, com os primeiros dados. A parte superior direita de cada gráfico mostra um número que representa o número de lances analisados.

Os gráficos da Fig. 2.2 mostram a evolução da distribuição conforme os dados vão sendo analisados. A posição do máximo vai sendo gradualmente deslocada, ao mesmo tempo em que a distribuição estreita-se em tormo do pico, mostrando que após uma grande quantidade de lances temos uma excelente estimativa de $H = 0.25$ para a moeda, baseado nos dados de uma simulação; então poderia-se pensar que os dados estejam se originando de uma "moeda tetraédrica"que possui uma face cara e três faces coroas.

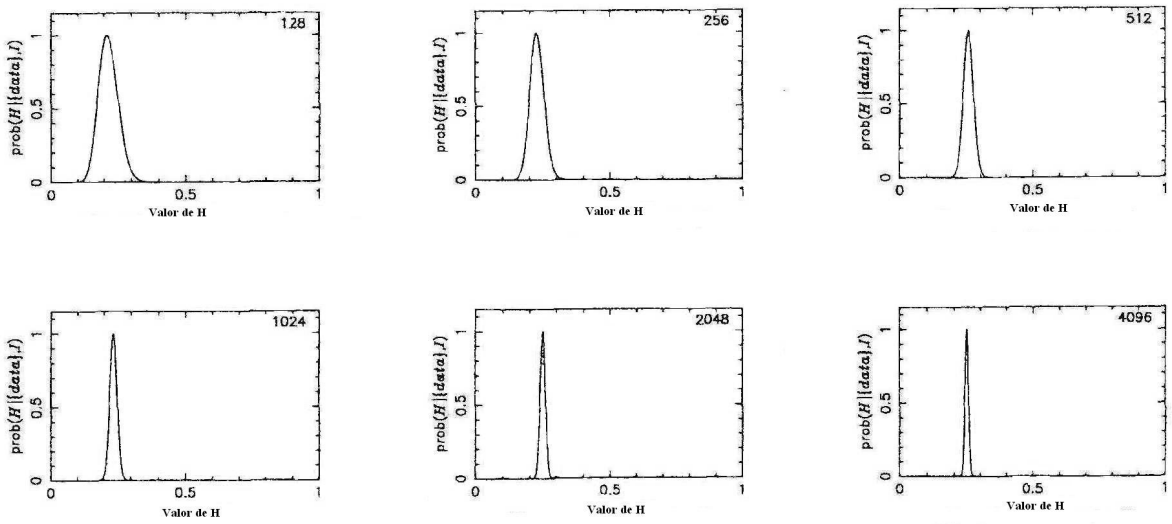

Figura 2.2: Evolução da distribuição posterior com os dados analisados. O pico vai estreitando-se no ponto H=0.25

2.2 Distribuições a Priori Diferentes.

A distribuição a priori uniforme (Eq. 2.11) foi escolhida por simplicidade; ela é o modo mais simples de expressar uma grande ignorância acerca da natureza da moeda. Uma questão interessante seria analisar como a inferência sobre a o parâmetro H da moeda varia quando escolhemos uma distribuição a priori diferente.

Para tratar esse problema, repetiremos a análise de dados anterior com duas distribuições a priori alternativas; os resultados são mostrados em (Fig. 2.3) e (Fig. 2.4)

A linha sólida representa a distribuição uniforme utilizada anteriormente, e está incluída com o objetivo de comparação. Uma das distribuições alternativas representada pela linha tracejada é um pico em torno de $H = 0.5$ refletindo que nossa informação inicial é de que esta seja uma moeda honesta, a largura desse pico é tal que permite que H seja tão pequeno quanto $H = 0.35$ ou tão alto quanto 0.65, mas com probabilidade muito baixa para tais resultados. A outra distribuição alternativa, representada por linhas pontilhadas, possui picos bastante estreitos em $H = 0$ e $H = 1$, o que indica esperarmos que a moeda seja altamente viciada, ou seja, produza como resultado apenas cara ou apenas coroa.

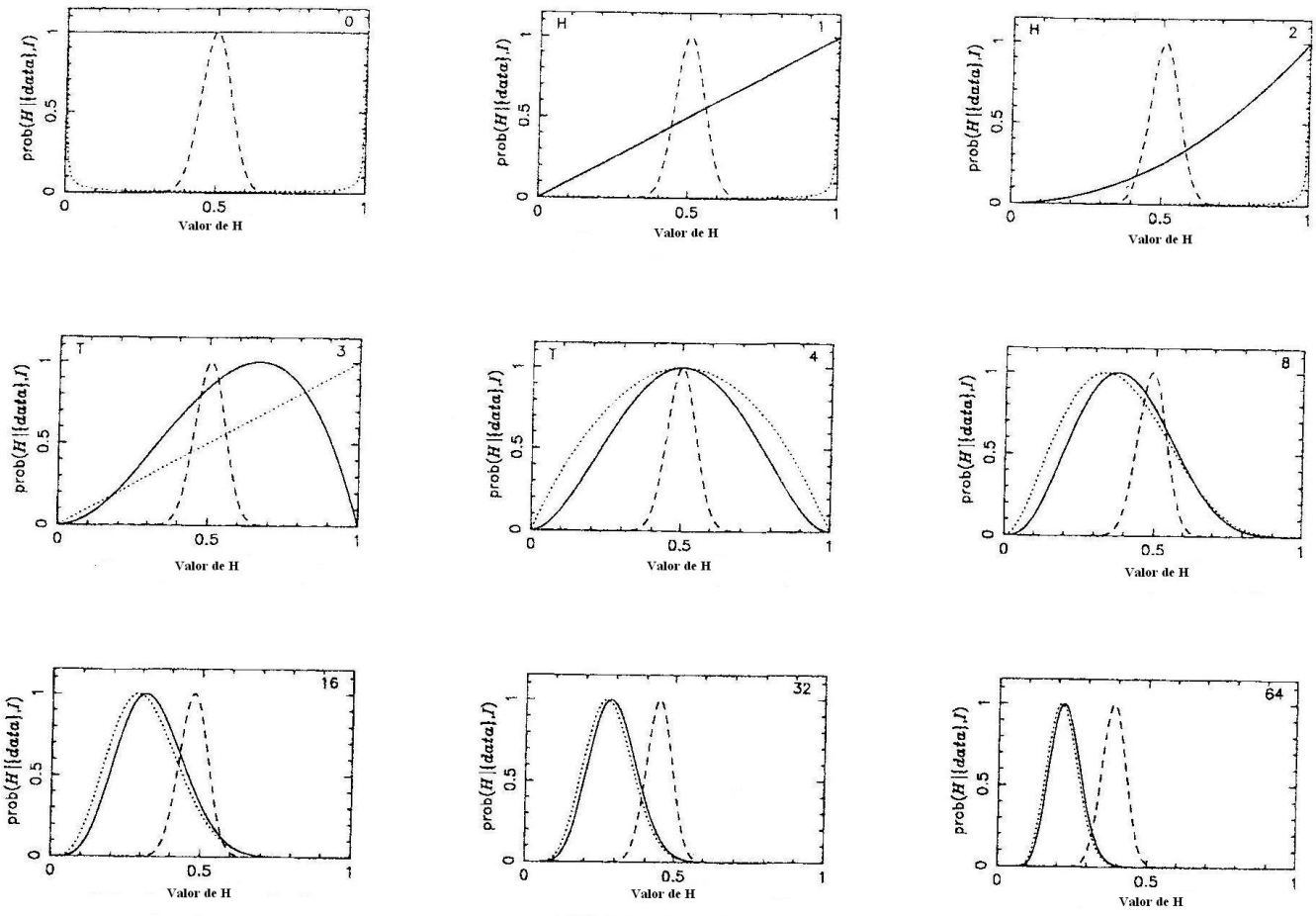

Figura 2.3: O efeito das diferentes distribuições a priori na evolução inicial da distribuição posterior do caso da moeda.

A Fig. 2.3 mostra a evolução das distribuições posteriores com os primeiros lances da moeda. Note que com poucos dados analisados as distribuições são notadamente diferentes entre si. Conforme o número de lances aumenta e mais dados são analisados, todos as distribuições tornam-se picos estreitos e acabam convergindo para um mesmo valor (ver Fig. 2.4).

A análise de poucos dados nos diz pouco sobre a natureza da moeda. Nosso grau de conhecimento depois da análise desses poucos dados é fortemente dependente do que sabíamos ou assumimos antes do resultado por isso, as distribuições posteriores são diferentes. No entanto, quando as evidências empíricas aumentam, eventualmente chegamos às mesmas conclusões independentemente de nossas suposições iniciais, a distribuição posterior é dominada então pela função de verossimilhança e a escolha da probabilidade a priori torna-se irrelevante. Note que, nesse caso, dispomos de um número

muito grande de dados, por isso, a priori não resultou numa influência muito grande no final da análise o que nem sempre é o caso.

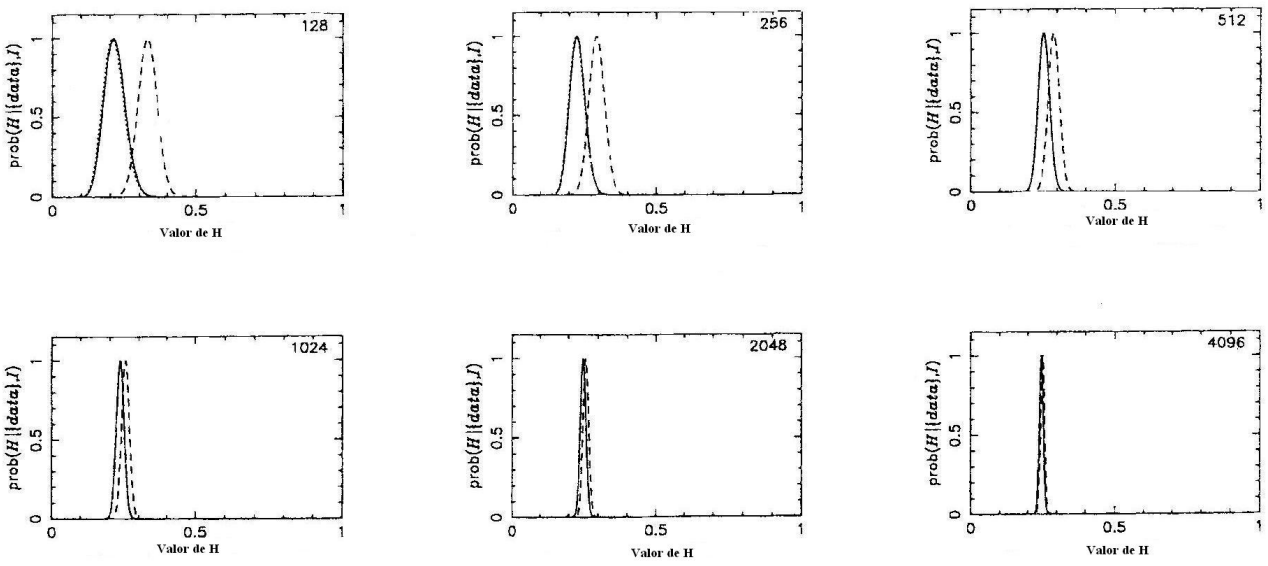

Figura 2.4: Evolução das várias distribuições posteriores construídas a partir de distribuições a priori diferentes. Note que todas as distribuições acabam convergindo

2.2.1 O Problema do Farol

O exemplo a seguir [17] consiste na seguinte questão: um farol localizado a uma distância α da costa e de altura β lança feixes de luz distribuídos aleatoriamente sobre a superfície da água, que são detectados nas diversas posições x_k, como ilustrado esquematicamente na figura 2.5 . A partir destes dados, desejamos determinar a posição do farol (α, β).

Dada a geometria do problema, é razoável supor que $p(\theta_k|\alpha, \beta, I)$, que é a probabilidade de detectar um feixe de luz em um determinado ângulo azimutal, seja independente do valor de θ_k. O valor dessa constante pode ser determinado levando-se em conta o fato de que os feixes são espalhados na superfície da água e, portanto, $-\pi/2 < \theta_k < \pi/2$. Assim, normalizando a função, obtemos:

$$p(\theta_k|\alpha, \beta, I) = \frac{1}{\pi} \tag{2.14}$$

Desde que os detectores de luz são sensíveis apenas à posição ao longo da costa e não à direção,

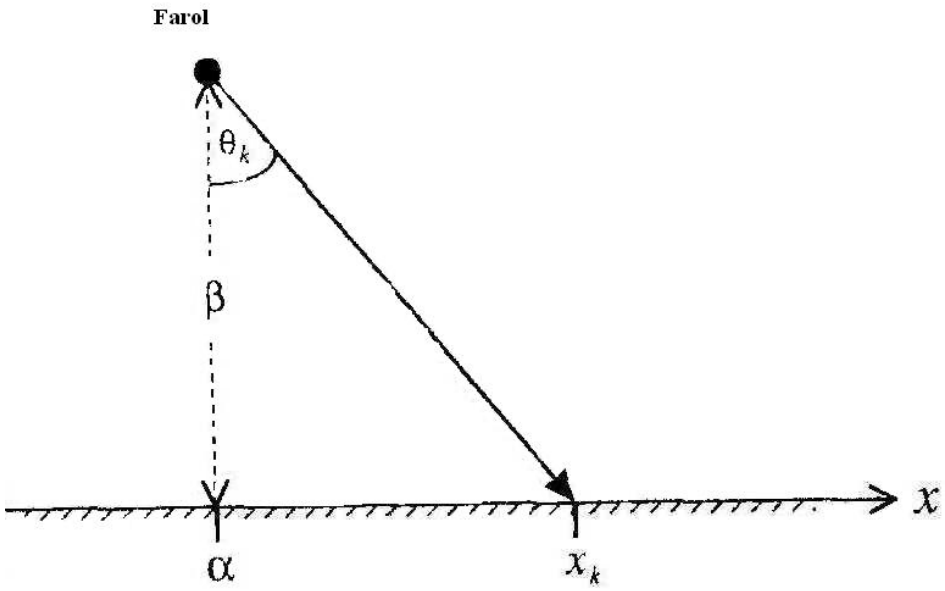

Figura 2.5: Ilustração esquemática da geometria do problema do farol.

devemos relacionar o ângulo azimutal θ_k com x_k. Observando a Fig. 2.5 e usando um pouco de geometria elementar:

$$\beta tan(\theta_k) = x_k - \alpha \tag{2.15}$$

Aplicando uma mudança de variável, podemos transformar a equação anterior em uma expressão de probabilidade que dependa do parâmetro de interesse x_k, usando a seguinte relação:

$$p(x_k|\alpha, \beta, I) = prob(\theta_k|\alpha, \beta, I)\frac{d\theta_k}{dx_k} \tag{2.16}$$

Desta forma, reescremos a distribuição de probabilidades Eq. (2.14), agora como função do parâmetro x_k:

$$p(x_k|\alpha, \beta, I) = \frac{\beta}{\pi[\beta^2 + (x_k - \alpha)^2]} \tag{2.17}$$

Esta distribuição de probabilidade é conhecida como *Lorentziana*. Ela é simétrica em torno do máximo, $x_k = \alpha$ e possui uma largura à meia altura (FWHM) de 2β

Com o objetivo de facilitar a análise do problema, consideraremos que β seja conhecido, reduzindo assim o problema em determinar a distância do farol em relação à costa. Para isso, vamos

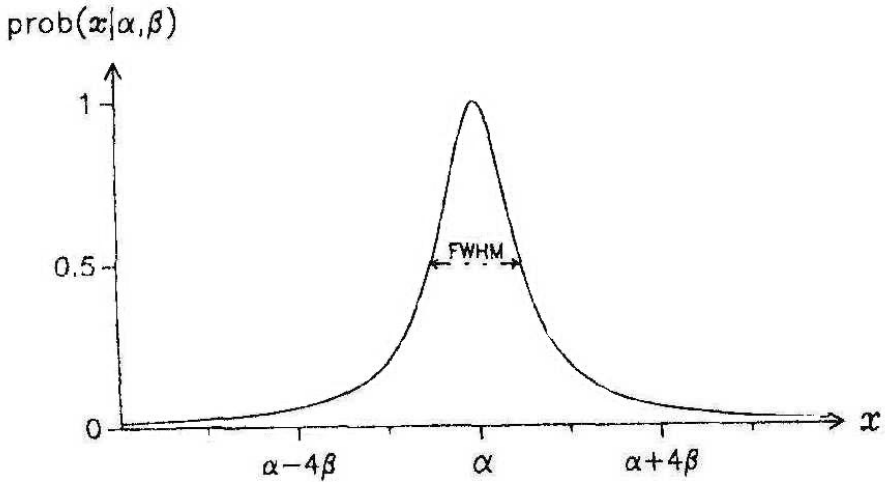

Figura 2.6: Distribuição Lorentziana ou de Cauchy, que é simétrica com respeito ao ponto de máximo.

recorrer ao Teorema de bayes que, nesse caso, toma a seguinte forma:

$$p(\alpha|x_k, \beta, I) \propto p(x_k|\alpha, \beta, I) \times p(\alpha|\beta, I) \tag{2.18}$$

Dado que qualquer conhecimento sobre o parâmetro β não interfere em nossa inferência sobre o valor de α a priori, podemos escrever:

$$p(\alpha|\beta, I) = p(\alpha|I) = A \tag{2.19}$$

Desde que a coleta de um sinal não influencia nas medidas posteriores, a função de verossimilhança (likelihood) para estes dados independentes é tão somente o produto das probabilidades de obter N detecções individuais:

$$p(\{x_k\}|\alpha, \beta, I) = \prod_{k=1}^{N} p(x_k|\alpha, \beta, I) \tag{2.20}$$

De posse das duas equações acima, procedemos substituindo-as no teorema de Bayes afim de obter a distribuição posterior ou, para facilitar os cálculos, o logaritmo dela:

$$L = ln\left[p(\alpha|\{x_k\}, \beta, I)\right] = constante - \sum_{k=1}^{N} ln(\beta^2 + (x_k - \alpha)^2) \tag{2.21}$$

onde a constante acima inclui termos que não envolvem α.

Para obtermos a melhor estimativa do valor de α, maximizamos a distribuição posterior, com relação à α:

$$\frac{dL}{d\alpha}|_{\alpha_0} = 2\sum_{k=1}^{N} \frac{x_k - \alpha_0}{\beta^2 + (x_k - \alpha_0)^2} = 0 \tag{2.22}$$

É complicado, no entanto, rearranjar os termos da equação de forma a expressar α_0 em função de x_k e β. Para contornar este problema, podemos utilizar uma análise numérica. Uma maneira bastante simples para isso é calcular para uma série de valores de α os respectivos valores de L e, posteriormente, exp(L). Para tanto, geramos uma distribuição uniforme de valores dos ângulos azimutais θ_k que são convertidos em valores de x_k a partir da Eq. 2.17. A melhor estimativa do parâmetro α corresponderá ao valor correspondente ao pico do gráfico.

Nas Fig. 2.7 e Fig. 2.8 é apresentado o resultado de uma simulação para a qual fixamos o valor de $\beta = 1$. Pode-se notar que para poucos valores gerados, o gráfico apresenta uma grande largura em torno do máximo (Fig. 2.7, primeiro painel) e até em alguns casos a presença de dois picos (o que significaria que o farol possui duas fontes) (Fig. 2.7, segundo painel).

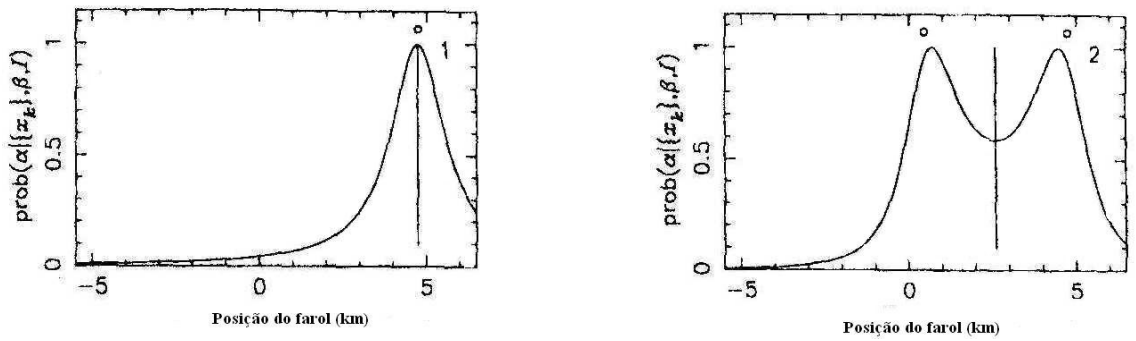

Figura 2.7: Distribuição posterior com respeito ao problema do farol, após a análise de respectivamente, um e dois dados. Os círculos pequenos em cima do gráfico marcam a posição em que foi recolhido cada flash de luz.

Conforme os valores de α vão evoluindo, a forma do gráfico da distribuição de probabilidade posterior vai estreitando-se continuamente até que, finalmente, acaba convergindo para $\alpha = 1$ (Fig. 2.8)

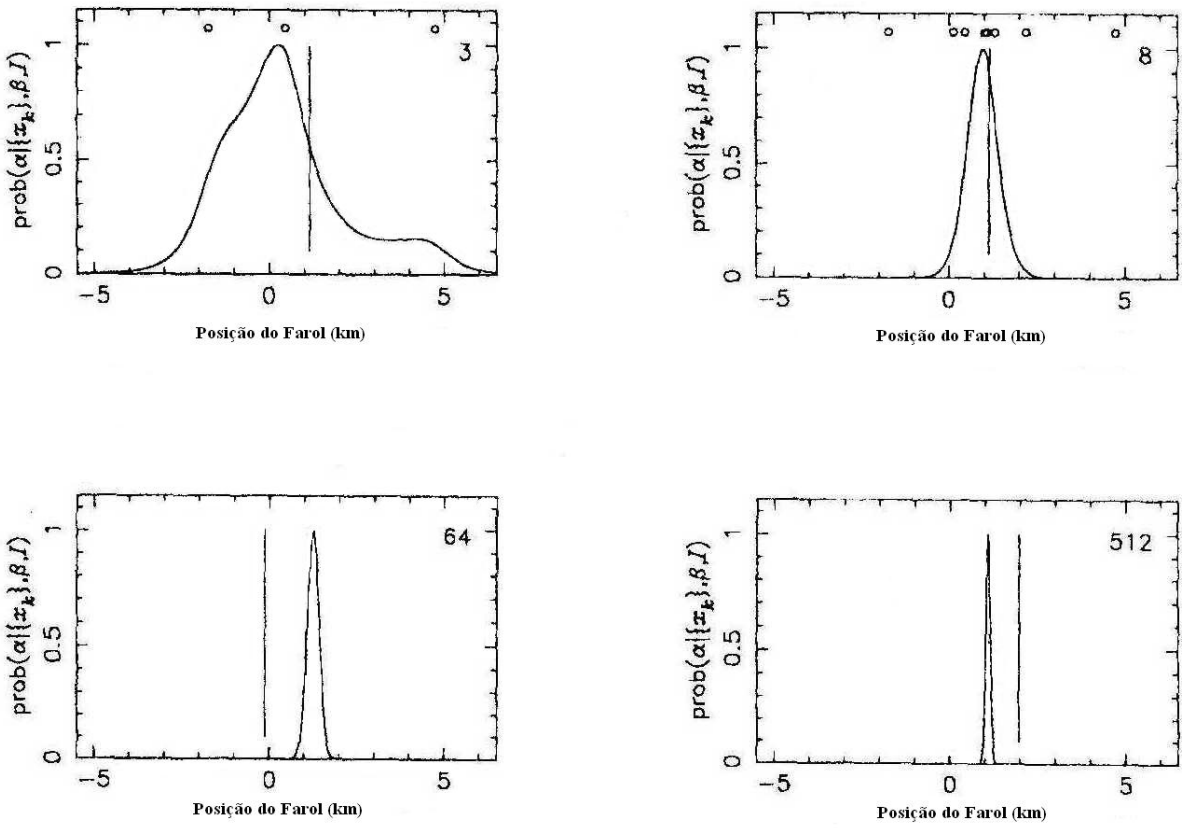

Figura 2.8: Evolução da distribuição posterior para a posição do farol com o número de dados avaliados, que é mostrado no na parte superior direita de cada gráfico. O traço vertical representa o valor médio dos dados.

2.3 Estimativa de vários parâmetros

Vamos generalizar nossa discussão para casos que envolvem a estimativa de mais de um parâmetro, usando como exemplo ilustrativo a determinação da amplitude de um sinal na presença de ruído no fundo. Os dados coletados são números inteiros que podem representar, por exemplo, o número de fótons a um certo comprimento de onda, ou o número de prótons espalhados em uma dada direção. Dado o conjunto de contagens N_k medidas em um conjunto experimental x_k, qual a melhor estimativa da amplitude do pico de sinal e do ruído de fundo?

Considerando que o pico tenha a forma de uma gaussiana, por exemplo, com largura ω e centrado em x_0, poderíamos atribuir para um determinado dado ideal D_k a seguinte expressão:

$$D_k = n_0 \left[Aexp\left(-\frac{(x_k - x_0)^2}{2\omega^2} \right) + B \right] \tag{2.23}$$

onde n_0 é uma constante relacionada a quantidade de tempo para a qual a medida foi feita.

Ao contrário do número de contagens N_k, D_k na Eq. 2.23 não é, em geral, um número inteiro. No entanto, o dado atual será dado por um inteiro na vizinhança de seu valor esperado; uma distribuição que incorpora essa propriedade e é usualmente usada em experimentos de contagem é a chamada distribuição de Poisson (ver Fig. 2.9):

$$p\,(N|D) = \frac{D^N e^{-D}}{N!} \tag{2.24}$$

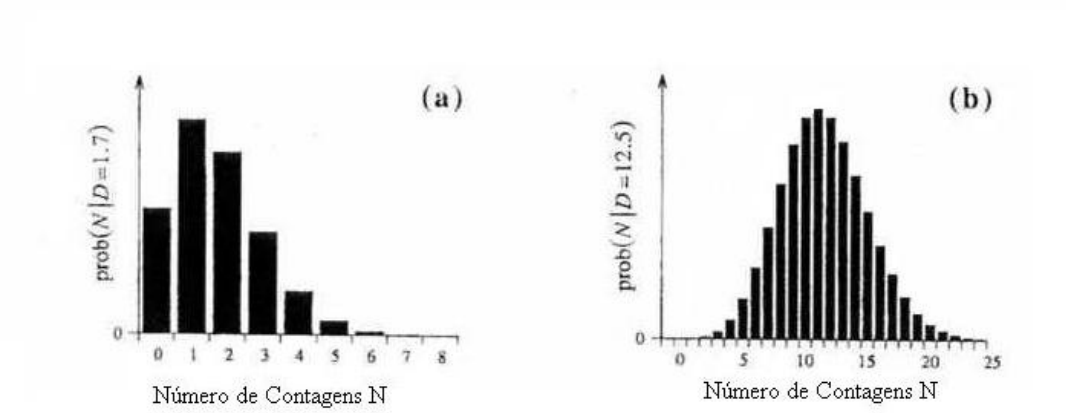

Figura 2.9: Exemplo de distribuições de Poisson, com $D = 1.7$ e $D = 12.5$, respectivamente.

Note que o valor esperado de N é dado por D:

$$\langle N \rangle = \sum_{N=0}^{\infty} N prob(N|D) = D \tag{2.25}$$

Portanto, podemos escrever como função de verosimilhança (likelihood function) do dado N_k a seguinte expressão:

$$p\,(N_k|A, B, I) = \frac{(D_k)^{N_k} e^{-D_k}}{N_k!} \tag{2.26}$$

onde a informação de *background* I incluiu nosso conhecimento sobre a relação entre o número esperado de contagens D_k e os parâmetros de interesse A e B.

Para o caso do modelo de pico gaussiano x_0, ω e n_0 são conhecidos, assim como x_k. Como os dados x_k são independentes entre si, de forma que dados os valores de A e B, o número de contagens observadas em um canal não influencia em qualquer outro canal, então a função likelihood é simplesmente o produto das medidas individuais:

$$p\left(\{N_k\}\,|A,B,I\right) = \prod_{k=1}^{M} p\left(N_k|A,B,I\right) \tag{2.27}$$

Usando o Teorema de Bayes:

$$p\left(A,B|\,\{N_k\}\,,I\right) \propto p\left(\{N_k\}\,|A,B,I\right) \times p\left(A,B|I\right) \tag{2.28}$$

Mais uma vez, vamos adotar uma grande ignorância inicial, e atribuiremos para a distribuição a priori uma constante, lembrando que $A \geq 0, B \geq 0$, pois amplitudes devem ter sinal positivo. Em caso contrário, a distribuição a priori toma o valor nulo. Desde modo, obtemos para a distribuição posterior:

$$p\left(A,B|N_k,I\right) \propto \prod_{k=1}^{M} \frac{(D_k)^{N_k} e^{-D_k}}{N_k!} \tag{2.29}$$

Tomando o ln da expressão acima obtemos, finalmente:

$$L = \ln P = constante + \sum_{k=1}^{M} N_k ln D_k - D_k \tag{2.30}$$

Nossa melhor estimativa para a amplitude do sinal e *background* é dado pelos valores de A e B que maximizam L. Quatro conjuntos de dados e respectivas distribuições posteriores são mostradas na Fig. 2.10. Os dados são representados em forma de histogramas, cada canal de dados com tamanho correspondendo a unidade. Em todos os casos, o sinal principal é centrado em $x_0 = 0$, tendo um FWHM de cinco unidades, onde se assume que esse valor seja conhecido em todos os casos. Como a distribuição nesse caso é bidimensional, pois depende de A e B é interessante o uso de contornos para representá-la, ou seja, linhas de igual densidade de probabilidade. Na Fig. (2.10, segunda coluna) os contornos correspondem a 10, 30, 50, 70 e 90 por cento de probabilidade máxima.

O primeiro painel mostra o número de contagens detectadas para 15 canais de dados, onde o parâmetro n_0 foi escolhido para ter um valor máximo de 100 contagens. A distribuição posterior

apresentada no segundo painel na parte superior indica que a melhor estimativa da amplitude do sinal é aproximadamente igual a unidade e cerca de metade do valor do ruído, B. Os painéis imediatamente abaixo correspondem ao mesmo conjunto experimental, mas com a coleta dos dados reduzida à um décimo do tempo; os dados aparecem mais espalhados e a distribuição posterior correspondente é cerca de três vezes mais larga que a anterior em ambas as direções, sendo truncada para valores negativos de A, o que reflete a importância da probabilidade a priori quando os dados são pobres.

No terceiro conjunto de dados, temos que a razão de contagem volta a seu valor inicial. No entanto, agora há 31 dados coletados em um intervalo experimental para os valores de x_k duas vezes maior. Com o dobro dos dados nossa espectativa é de que a estimativa sobre os valores de A e B seja melhorada por um fator de $\sqrt{2}$, contudo, é difícil afirmar isso à partir dos diagramas; isso parece ser verdade apenas para o *background*, devido ao fato de que medidas muito longe da origem nada nos dizem sobre o pico do sinal mas contribuem somente para a inferência do valor de *background*.

O último conjunto de dados ilustra o caso em que há apenas sete dados contidos num intervalo experimental que é a metade do original. A distribuição é bastante larga, isso indica uma forte correlação entre nossas estimativas de A e B; como o intervalo dos x_k sobre os quais os dados são coletados é muito pequeno torna-se dificíl distinguir o sinal do ruído.

2.4 Distribuições Marginais

Gostaríamos de estimar somente a amplitude do sinal independentemente do ruído, ou seja, queremos determinar $p\left(A|\left\{N_k\right\},I\right)$. Para tanto, podemos reescrever a última expressão como sendo uma integral sobre todos os valores de B, da seguinte forma:

$$p\left(A|\left\{N_k\right\},I\right)=\int_0^\infty p\left(A,B|\left\{N_k\right\},I\right)dB \tag{2.31}$$

Ou, se quisermos o oposto:

$$p\left(B|\left\{N_k\right\},I\right)=\int_0^\infty p\left(A,B|\left\{N_k\right\},I\right)dA \tag{2.32}$$

Assim, a análise dos dados é mais fácil de ser realizada, como pode ser visto, a partir da Fig. 2.11.

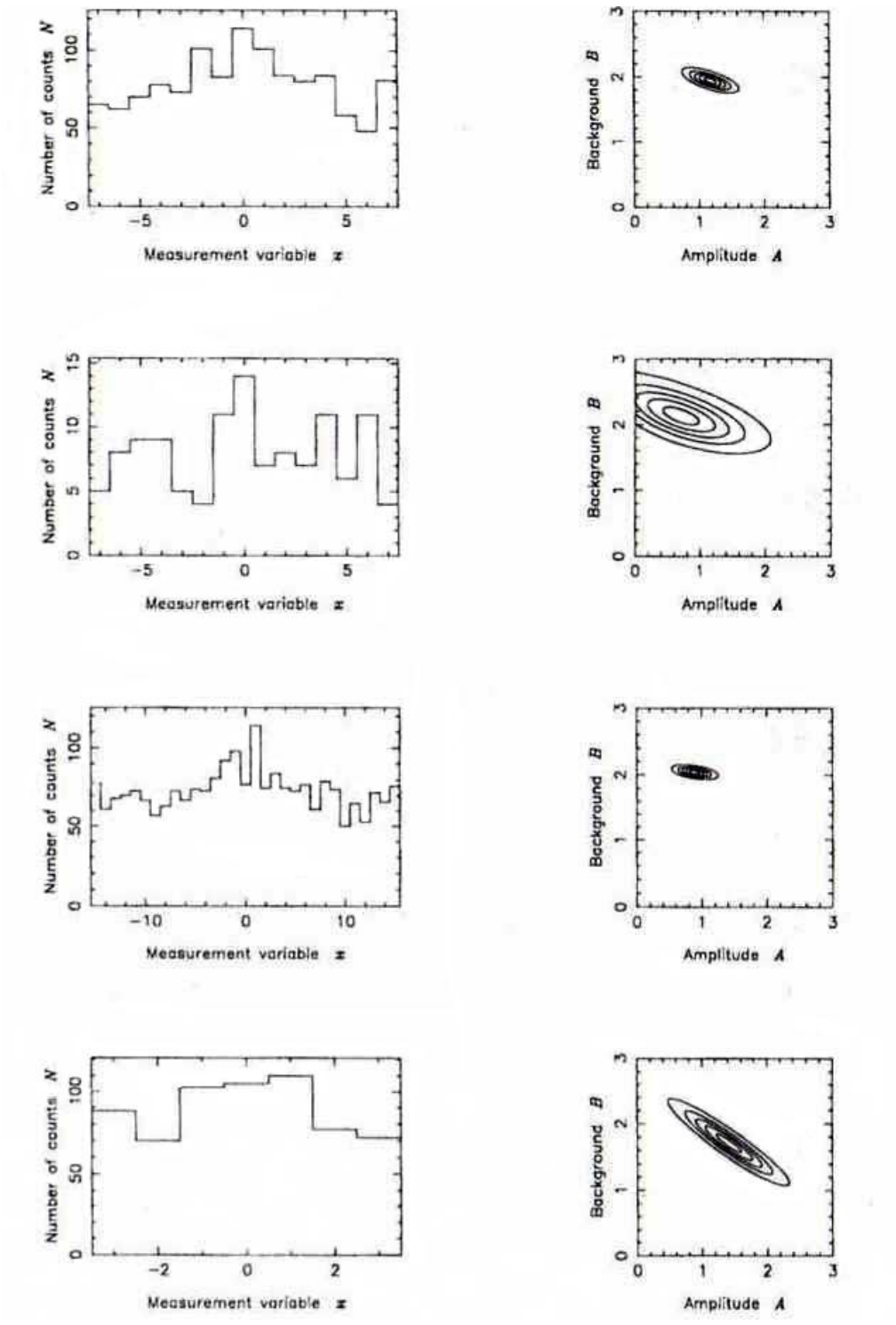

Figura 2.10: Dados de Poisson e as distribuições posteriores resultantes para a amplitude A de um pico gaussiano, e um ruído de fundo B, a partir de quatro diferentes arranjos experimentais.

Dessa forma, estamos marginalizando nossa distribuição posterior originalmente bidimensional, em duas distribuições cada qual dependendo de um único fator, A ou B.

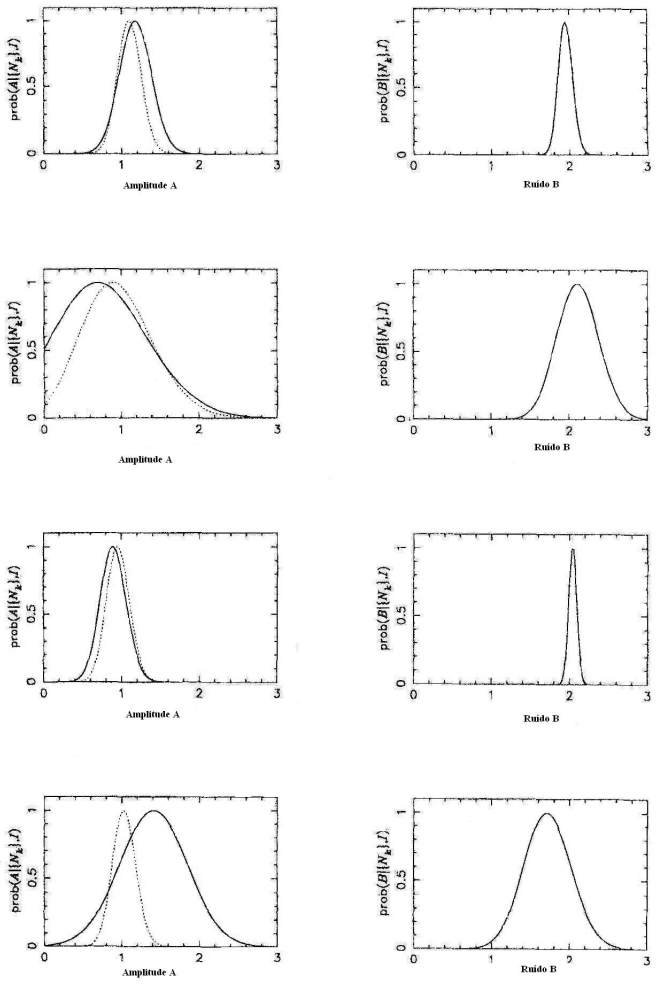

Figura 2.11: Distribuições marginalizadas para a amplitude A e o ruído B correspondentes ao arranjo experimental mostrado na figura anterior. A linha pontilhada representa a distribuição posterior de A condicional ao conhecimento de verdadeiro valor de B, $p(A|\{N_k\}, B, I)$.

Deve-se notar a diferença entre $p(A|\{N_k\}, I)$ e $p(A|\{N_k\}, B, I)$. A primeira leva em conta o fato de desconhecermos o valor de B. A última é apropriada quando já possuímos conhecimento prévio do valor de B. O gráfico da Fig. 2.11 mostra a diferença entre os dois. Pode-se notar que a curva da distribução marginal é mais larga que a da distribuição condicional; isto acontece pela diferença na quantidade de informação envolvida no problema, que é maior no caso da última, pois há o conhecimento do parâmetro B.

Em toda a análise feita anteriormente partimos do fato de que conhecíamos os valores de ω e x_0 do modelo gaussiano. E se essa condição for relaxada? Podemos usar a marginalização desses parâmetros:

$$p\left(A, B|N_k, I\right) = \int\int p\left(A, B, \omega, x_0|N_k, I\right) d\omega dx_0 \tag{2.33}$$

Note que, usando o Teorema de Bayes:

$$p\left(A, B, \omega, x_0|N_k, I\right) \propto p\left(N_k|A, B, \omega, x_0, I\right) \times p\left(A, B, \omega, x_0|I\right) \tag{2.34}$$

O primeiro termo do lado direito é equivalente a função de verossimlhança das Eq. (2.26) e (2.27), enquanto o segundo termo é a probabilidade a priori para os parâmetros A, B, ω e x_0. Esta pode ser decomposta pela regra do produto em:

$$p(A, B, \omega, x_0|I) = p(A, B|I) \times p(\omega, x_0|I) \tag{2.35}$$

Se já conhecemos a largura e a posição do pico gaussiano, podemos representar as distribuições a priori de ω, x_0 como sendo funções delta:

$$p(x_0, \omega|I) = \delta(\omega - \omega_0)\delta(x_0) \tag{2.36}$$

E, neste caso, a integral da (Eq. 2.33) é fácil de ser calculada e vale:

$$p(A, B|\left\{N_k\right\}, I) \propto p(\left\{N_k\right\}|A, B, \omega = \omega_0, x_0 = 0) \times p(A, B|I) \tag{2.37}$$

Como esperado, recuperamos a equação inicial do problema, quando os valores dos parâmetros eram conhecidos. No entanto, se não tivéssemos conhecimento de tais valores, poderíamos atribuir distribuições a priori para cada um dos parâmetros e calcular novamente as integrais.

2.5 Cálculos envolvendo a aplicação do método de Bayes

Apresentamos a seguir uma contribuição nossa no estudo de inferência paramétrica, baseado em uma *distribuição gama*. Ela aparece de várias formas como solução em equações de transporte de raios cósmicos e também em perda de energia de partículas carregadas a baixas energias. Adotamos

um procedimento similar a alguns exemplos apresentados anteriormente. Geramos eventos a partir de uma simulação Monte Carlo e posteriormente aplicamos o teorema de Bayes para estimativa dos parâmetros presentes na distribuição.

2.5.1 Metodologia

Justifica-se o uso da técnica de Monte Carlo em situações quando esta é frequentemente o único método prático de resolver equações de transporte, nas quais são geradas amostras de variáveis randômicas governadas por funções de densidade de probabilidade complicadas. Em particular, no nosso trabalho, usamos a técnica para gerar variáveis oriundas principalmente da distribuição Gama. A técnica de Monte Carlo assume o uso de um "gerador de variáveis randômicas"que gera valores uniformes estatisticamente independentes em um intervalo semi-aberto $[0, 1)$. A partir de tal distribuição uniforme podemos fazer uso de várias técnicas para transformar tal distribuição naquela desejada. Apresentaremos a seguir, a técnica que foi utilizada por nós para a obtenção da distribuição Gama.

Dada uma função de densidade de probabilidade $f(x)$, onde $-\infty < x < \infty$, define-se a distribuição de probabilidades cumulativa como sendo a probabilidade de que x assuma uma série de valores até um determinado limite, por exemplo, $x \leq a$:

$$F(a) = \int_{-\infty}^{a} f(x)dx \tag{2.38}$$

Se a é escolhido com a densidade de probabilidade $f(a)$, então a probabilidade integrada até a, $F(a)$, é por si mesma, uma variável aleatória que ocorrerá com densidade de probabilidade uniforme em $[0, 1)$. Se x pode assumir qualquer valor, e ignorando os extremos, podemos encontrar um único x escolhido a partir da distribuição $F(a)$ para um dado u, se tomarmos:

$$u = F(x) \tag{2.39}$$

contanto que possamos encontrar uma inversa de F, definida por:

$$x = F^{-1}(u) \tag{2.40}$$

Este método é mais apropriado quando pode-se calcular analiticamente a inversa da função da

integral indefinida de f.

2.5.2 Algoritmo que gera a distribuição gama

A forma analítica da função gama é dada por:

$$p(x|\lambda, k) = \frac{x^{k-1}\lambda^k e^{-\lambda x}}{\Gamma(k)} \tag{2.41}$$

O algoritmo que segue é dado para $\lambda = 1$. Para $\lambda \neq 1$ deve-se dividir o valor aleatório x encontrado por λ. Podemos dividir o algoritmo em três partes:

- Se k=1, temos uma distribuição exponencial. A partir do número aleatório uniforme gerado u, usa-se simplesmente a relação $x = -\ln u$

- Se $0 < k < 1$, inicia-se com $v_1 = (e + k)/e$, onde e é a base do logaritmo natural. Gera-se outros valores u_1, u_2 e define-se a quantidade $v_2 = v_1 u_1$. A partir de então, temos dois casos:

  - Caso 1: $v_2 \leq 1$. Define-se $x = v_2^k$ e verifica-se se $u_2 < e^{-x}$, em cujo caso aceitamos x, ou, reiniciamos gerando novos valores de u_1 e u_2

  - Caso 2: $v_2 > 1$. Define-se $x = -\ln\left([v_1 - v_2]/k\right)$. Se $u_2 < x^{k-1}$, então aceita-se x e o algoritmo é encerrado. Caso contrário, deve-se gerar novos u_1 e u_2.

- Se $k > 1$, inicia-se definindo $c = 3k - 0.75$; em seguida gerando u_1 e computando o valor $v_1 = u_1(1 - u_1)$ e $v_2 = (u_1 - 0.5)\sqrt{c/v_1}$. Se $x = k + v_2 - 1 \leq 0$, é necessário voltar e gerar novo u_1; caso contrário, gera-se u_2 e define-se $v_3 = 64v_1^3u_2^2$. Se $v_3 \leq 1 - 2v_2^2/x$ ou $_3 \leq 2[k-1]\ln[x/(k-1)] - v_2$ aceita-se x, caso contrário é necessário gerar novo u_1.

2.5.3 Aplicação do Teorema de Bayes

De posse dos eventos gerados pela simulação, estamos em condições agora de proceder o cálculo escrevendo o teorema de Bayes:

$$p(\lambda, k|\{x_i\}, I) \propto p(\{x_i\}|\lambda, k, I) \times p(\lambda, k|I) \tag{2.42}$$

Atribuiremos para a distribuição a priori uma constante, com a condição de que $\lambda > 0$ e $k > 0$. Para a função de verossimilhança, tomaremos simplesmente como a produtória da Eq.2.41:

$$p(\{x_i\}|\lambda, k, I) = \prod_{i=1}^{N} p(x_i|\lambda, k, I) = \left[\frac{\lambda^k}{\Gamma(k)}\right]^N \left[\prod_{i=1}^{N} x_i^{k-1}\right] exp(-\lambda \sum_{i=1}^{N} x_i) \tag{2.43}$$

Queremos agora marginalizar a distribuição a fim de obter o valor mais provável de cada parâmetro separadamente. Iniciaremos marginalizando o parâmetro λ para, assim, conseguir uma expressão para a distribuição posterior que seja dependente somente de k. Dessa forma, aplicamos a regra de marginalização:

$$p(k|\{x_i\}, I) = \int_0^\infty p(\lambda, k|\{x_i\}, I) d\lambda \tag{2.44}$$

Assim, calculando a integral:

$$p(k|\{x_i\}, I) = \left[\prod_{i=1}^{N} x_i^{k-1}\right] \frac{1}{\Gamma(k)^N} \int_0^\infty \lambda^{kN} exp(-\lambda \sum_{i=1} x_i) d\lambda \tag{2.45}$$

$$p(k|\{x_i\}, I) = \frac{1}{\left[\sum_{i=1}^{N} x_i\right]^{kN+1}} \frac{\Gamma(kN+1)}{[\Gamma(k)]^N} \left[\prod_{i=1}^{N} x_i^{k-1}\right] \tag{2.46}$$

Tomando o $\ln p(k|\{x_i\}, I)$, temos:

$$\ln p = \ln \Gamma(kN+1) - N \ln \Gamma(k) - (kN+1) \ln \sum_{i=1}^{N} x_i + (k-1) \sum_{i=1}^{N} \ln x_i \tag{2.47}$$

Calculamos agora para uma série de valores de k os respectivos valores de $\ln P$ e, posteriormente, P, de modo a obter o valor que maximize essa expressão. Todas as simulações foram feitas usando como parâmetros fixos $\lambda = 10.0$ e $k = 5.0$. Os gráficos da Fig. 2.12 mostram a evolução da distribuição posterior do parâmetro λ com o número de eventos. Na tabela 2.1 mostramos, para cada simulação, o valor mais provável do parâmetro k e a respectiva largura em torno deste máximo.

Inicialmente, analisamos uma simulação com 10 eventos. Note que o gráfico (ver Fig. 2.12) tem uma largura considerável em torno de seu máximo, indicando que há um número insuficiente de eventos para uma análise mais apurada. Na Fig. 2.13 traçamos o gráfico da distribuição calculada analiticamente a partir do resultado dos máximos obtidos pela marginalização e, juntamente com este, desenhamos o histograma dos eventos.

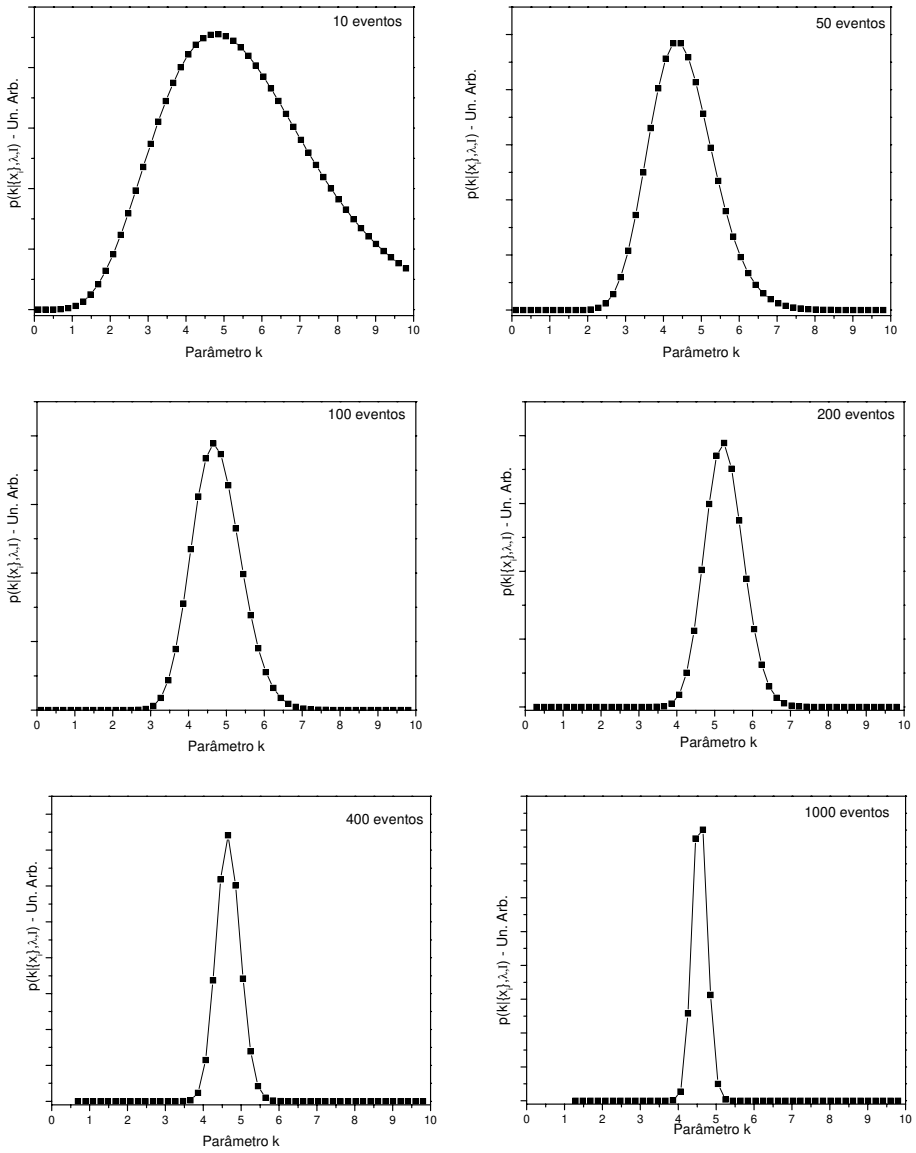

Figura 2.12: Evolução da distribuição posterior para o caso do modelo gama, na qual marginalizamos o parâmetro λ e estimamos k, a partir de um número reduzido de eventos (dez), aumentando progressivamente. Note que a largura do gráfico vai estreitando, indicando uma boa confidência na estimativa do parâmetro.

Tabela 2.1: Valores máximos de k e as respectivas larguras em torno destes.

Número de eventos simulados	**Valor máximo de k**	**Largura**
10	4.852	4.752
50	4.456	1.98
100	4.515	1.584
200	5.248	1.188
400	4.654	0.594
1000	4.654	0.396

Para a marginalização do parâmetro k utilizamos um procedimento diferente do anterior. Ao invés do cálculo da integral sobre todos os valores possíveis de k, nos valemos de uma propriedade da distribuição gama, que é dada por:

$$\langle x \rangle = \frac{k}{\lambda} \tag{2.48}$$

onde tomamos $\langle x \rangle$ como um valor conhecido.

Assim, substituindo k por $\lambda \langle x \rangle$ na distribuição posterior teremos uma expressão que só dependa de λ:

$$P = p(\lambda|k, \{x_i\}, I) = \frac{1}{\Gamma(\lambda \langle x \rangle)^N} \lambda^{\lambda \langle x \rangle N} \prod_{i=1}^{N} x_i^{\lambda \langle x \rangle - 1} e^{-\lambda \langle x \rangle N} \tag{2.49}$$

Tomando o $\ln P$:

$$\ln P = \lambda \langle x \rangle N ln\lambda - N ln\Gamma(\lambda \langle x \rangle) + (\lambda \langle x \rangle - 1) \sum_{i=1}^{N} lnx_i - N\lambda \langle x \rangle \tag{2.50}$$

De posse da expressão para o $\ln P$, repetimos os cálculos que fizemos anteriormente, agora para este caso. Pode ser visto pelos gráficos da evolução posterior do parâmetro λ com o número de eventos analisados (Fig. 2.14), bem como pelo gráfico em que se compara a função calculada analiticamente com os histogramas (Fig. 2.15) que esta análise também obteve para um conjunto

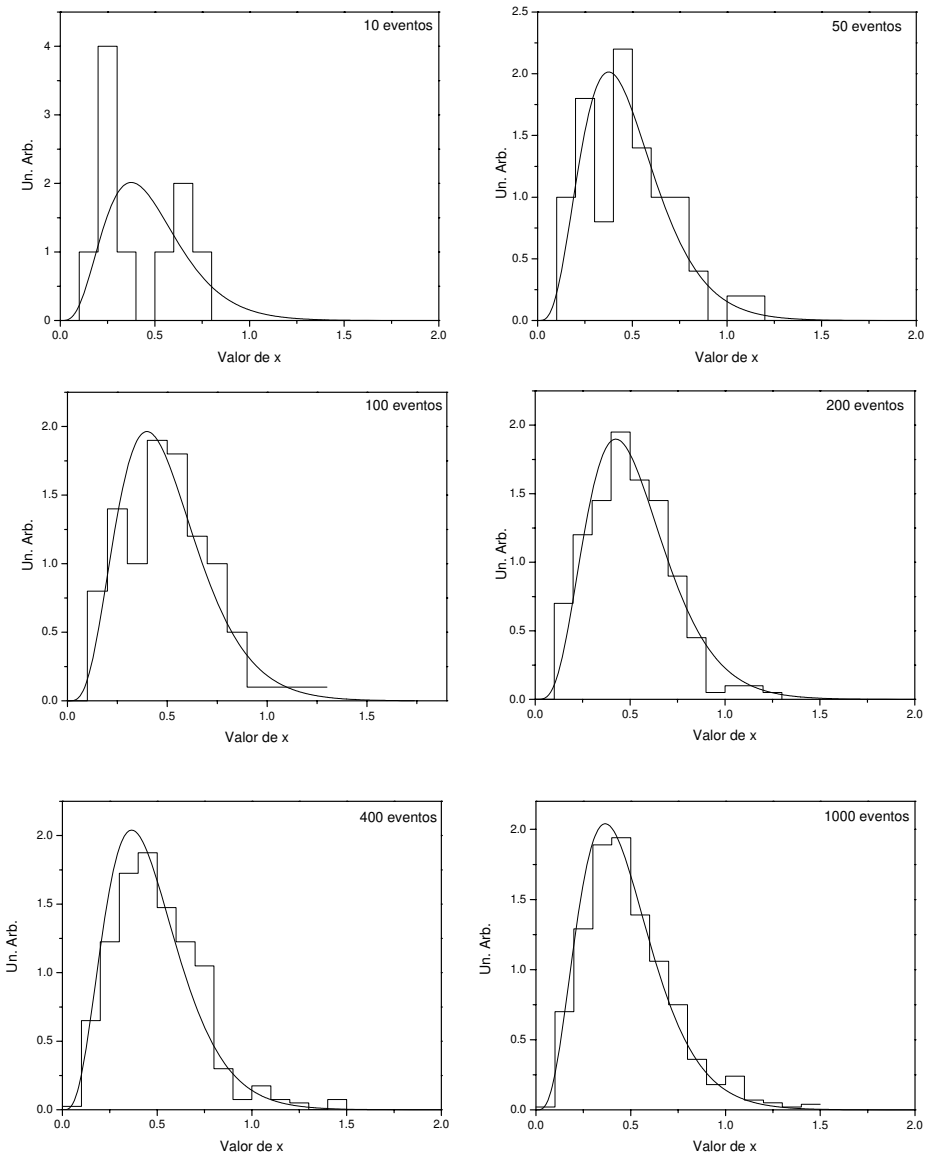

Figura 2.13: Comparação entre os histogramas gerados e a distribuição gama calculada com os parâmetros obtidos pelo aplicação do método de Bayes

razoável de eventos uma boa estimativa para o parâmetro λ, chegando muito próximo do valor de entrada da simulação ($\lambda = 10.0$). Desse modo, acreditamos ter obtido bons resultados numéricos

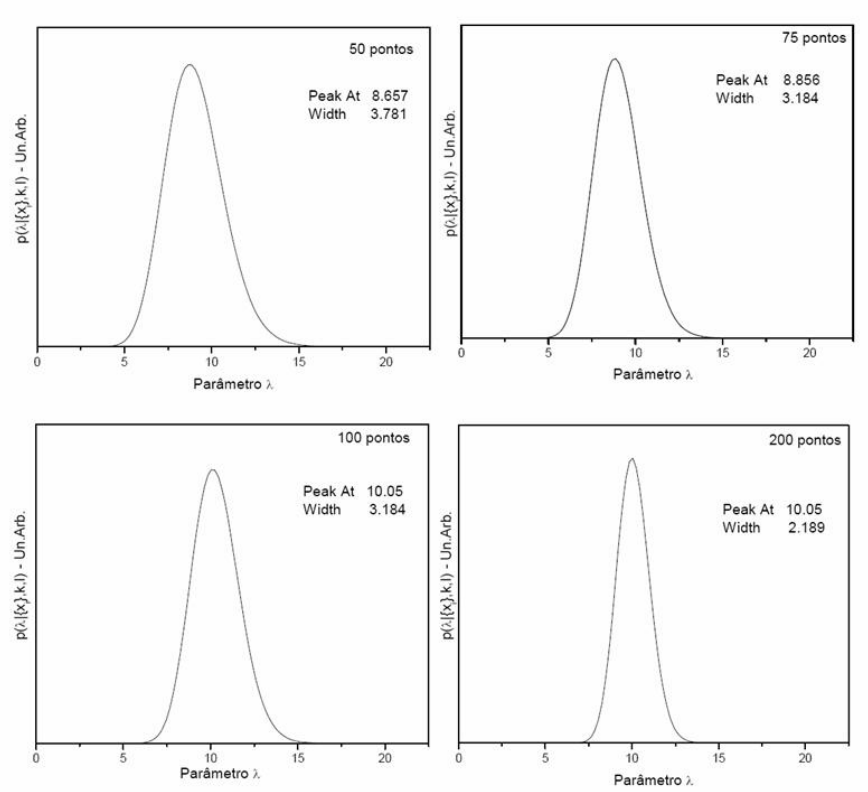

Figura 2.14: Evolução da distribuição posterior do parâmetro com os dados. Note que a diferença de estimativa para 100 e 200 eventos não é alterada, no entanto, a largura vai estreitando cada vez mais.

no uso do método bayesiano em uma aplicação de inferência paramétrica.

2.6 Escolha da Probabilidade a Priori

Na última seção consideramos situações dominadas principalmente pelas funções de verossimilhança, em que a distribuição a priori poderia ser incluída apenas como uma constante de normalização. No entanto, devemos ficar atentos à possibilidade do uso não-crítico de probabilidades a priori uniformes como uma prescrição ou uma regra, quando de fato o uso da distribuição a priori está intimamente relacionado com o tipo de informação disponível, que varia caso-a-caso.

Como pode-se imaginar, a escolha da priori é um assunto altamente discutido entre aqueles que se utilizam dos métodos bayesianos. O objetivo desta seção é tão somente o de mostrar alguns argumentos normalmente usados na escolha das distribuições a priori, bem como citar algumas

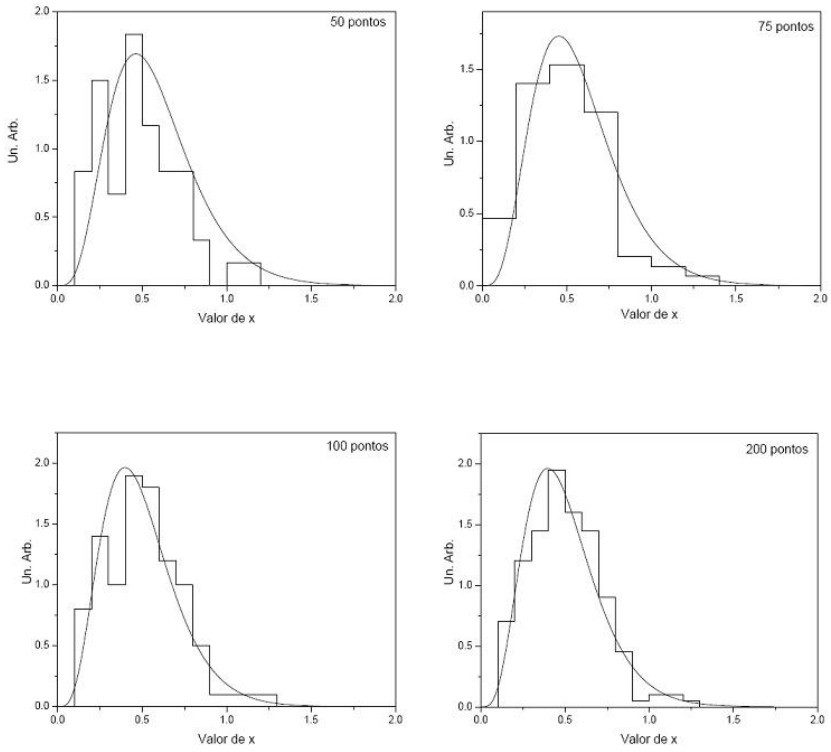

Figura 2.15: Comparação dos histogramas gerados a partir da simulação e a distribuição gama com os parâmetros obtidos a partir dos cálculos de marginalização

probabilidades a priori conhecidas.

2.6.1 Distribuições a Priori Conjugadas

Devido a problemas computacionais a modelagem das probabilidades a priori tem sido tradicionalmente uma escolha entre aquela que seria a mais honesta possível com a realidade e que ao mesmo tempo fosse uma função matemática que simplifique os cálculos analíticos. Uma estratégia bastante conhecida é a de escolher uma distribuição a priori com uma forma conveniente tal que a distribuição posterior pertença à mesma família funcional da distribuição a priori. A escolha da distribuição a priori depende então também da verossimilhança; desta forma dizemos que a distribuição a priori e a posterior estão *conjugadas*. Um caso interessante é obtido pela verossimilhança binomial $P(n|\theta, N)$, que tem uma forma proporcional a $\theta^n(1-\theta)^{N-n}$, apresentando deste modo a mesma estrutura da distribuição Beta:

$$f(\theta|r,s) = \frac{1}{\beta(r,s)}\theta^{r-1}(1-\theta)^{s-1} \tag{2.51}$$

onde $0 \leq \theta \leq 1$, $r, s > 0$ e $\beta(r,s)$ é a função Beta, definida como:

$$\beta(r,s) = \int_0^1 \theta^{r-1}(1-\theta)^{s-1}d\theta \tag{2.52}$$

Dependendo do valor dos parâmetros, a distribuição Beta pode tomar uma grande variedade de formas. Por exemplo, para grandes valores de r e s a função é muito similar a distribuição gaussiana, enquanto que para $r = s = 1$ uma distribuição constante é obtida.

Usando a distribuição Beta como priori em problemas de inferência com uma verossimilhança binomial, temos:

$$P(\theta|n,N,r,s) \propto [\theta^n(1-\theta)^{N-n}][\theta^{r-1}(1-\theta)^{s-1}] \propto \theta^{n+r-1}(1-\theta)^{N-n+s-1} \tag{2.53}$$

A distribuição posterior ainda é uma Beta com $n' = r+n$ e $s' = s+N-n$. Note que a posterior torna-se progressivamente independente da informação a priori no limite em que muitas dados são avaliados; neste limite obtemos o mesmo resultado que uma priori uniforme ($r = s = 1$)

Tabela 2.2: Algumas distribuições a priori conjugadas; x e n representam os valores observados (respectivamente contínuos e discretos) e θ é o parâmetro que se deseja inferir, correspondendo ao μ de uma Gaussiana, θ de uma binomial e λ de uma distribuição de Poisson.

Verossimilhança $P(x\|\theta)$	**Priori conjugada** $P(\theta)$	**Posterior** $P(\theta\|x)$
Normal(θ, σ)	Normal(μ_0, σ_0)	Normal(μ_1, σ_1)
Binomial(N, θ)	Beta(r,s)	Beta$(r+n, s+N-n)$
Poisson(θ)	Gamma(r, s)	Gamma$(r+n, s+1)$
Multinomial$(\theta_1, ..., \theta_k)$	Dirichlet$(\alpha_1, ..., \alpha_k)$	Dirichlet$(\alpha_1+n_1, ..., \alpha_k+n_k)$

2.6.2 Invariância por Transformações

Uma importante classe de prioris surgem como resultado de uma invariância por transformações. Consideraremos aqui dois casos: invariância por translação, invariância por escala.

Invariância por Translação

Assumindo que há indiferença quanto a transformações do tipo $\theta' = \theta + b$, onde θ é nossa variável de interesse e b uma constante. A invariância por translação requer que o elemento infinitesimal de probabilidade de que θ esteja no intervalo $d\theta$, $p(\theta)d\theta$ mantenha-se inalterado quando expressado em termos de θ', ou seja:

$$p(\theta)d\theta = p(\theta')d\theta' = p(\theta + b)d\theta \tag{2.54}$$

Para que a equação anterior seja válida para qualquer b, $p(\theta)$ deve ser igual a uma constante para todos os valores de θde $-\infty$ até ∞. Na prática, esta priori pode ser considerada como sendo $p(\theta) = 1/\Delta\theta$, onde $\Delta\theta$ é um intervalo finito muito grande ao redor dos valores de interesse.

Invariância por escala

Em outros casos, poderíamos estar indiferentes quanto uma transformação de escala, isto é, $\theta' = \beta\theta$, onde β é uma constante. Como $d\theta' = \beta\theta$, temos que:

$$P(\theta)d\theta = P(\beta\theta)\beta\theta \tag{2.55}$$

$$P(\beta\theta) = \frac{P(\beta)}{\beta} \tag{2.56}$$

cuja solução é dada por $P(\theta) \propto \frac{1}{\theta}$, $0 < \theta < \infty$.

Esta é conhecida como a função a priori de Jeffreys [10]. Note que esta priori pode ser escrita como $P(ln\theta) = cte$. Os valores de θ foram restringidos a apenas os positivos porque tradicionalmente variáveis que satisfazem esta invariância são associados com quantidades definidas positivamente. Variáveis associadas à invariância de translação são chamadas de *parâmetros de locação*, como o parâmetro μ no modelo gaussiano; valores associados à invariância por escala são chamados de *parâmetros de escala*, como σ no modelo gaussiano ou λ no modelo de Poisson. Esta é uma priori

não-informativa muito útil em muitas aplicações importantes e que reflete completa ignorância a priori, pode-se perceber o uso dela nos exemplos anteriores, onde se discutia o uso de análise de dados e as escolhas de modelos de verossimilhanças.

2.7 Seleção de Modelos

Até então temos considerado casos envolvendo estimativa de parâmetros e escolha de distribuições a priori. Vamos agora analisar casos em que há incerteza quanto ao modelo a ser usado. Inicialmente, poderíamos pensar que uma escolha entre diferentes alternativas propostas pode ser feita apenas com base em quão bem tais modelos reproduzem o resultado. No entanto, alguma reflexão revela a potencial dificuldade em que modelos mais complicados, definidos por muitos parâmetros, sempre serão capazes de estar em melhor acordo com as medidas experimentais, no entanto, deve-se levar em conta que tais são muito mais complexos do ponto de vista algébrico, por isso deve haver um ponderamento entre esses fatores. Apresentaremos a seguir, uma discussão sobre essas questões baseadas em um formulação elementar devida a Jeffreys, chamada de a história do Sr.A e do Sr.B. A questão que se apresenta é a seguinte:

O Sr. A tem uma teoria; O Sr. B também tem uma teoria, mas com um parâmetro ajustável λ. Quais das teorias é preferível com base nos dados D?

É evidente que, a partir das discussões precedentes, é necessário calcular as probabilidades posteriores que A e B estejam corretas e compará-las, assim:

$$\sigma = \frac{P(A|D,I)}{P(B|D,I)} \tag{2.57}$$

Deste modo, se $\sigma >> 1$ então preferiremos a teoria A, se $\sigma << 1$ preferiremos a teoria B, e caso $\sigma \approx 1$ não temos um conjunto de dados D suficientes para fazer uma escolha. Para calcular esta razão, vamos aplicar o teorema de Bayes tanto no numerador quanto no denominador:

$$\frac{P(A|D,I)}{P(B|D,I)} = \frac{P(D|A,I)}{P(D|B,I)} \times \frac{P(A|I)}{P(B|I)} \tag{2.58}$$

Faremos uma suposição que as razões entre as distribuição a priori seja igual a unidade. Para atribuir as probabilidades envolvendo as verossimilhanças $P(D|A,I)$ e $P(D|B,I)$ devemos conseguir

comparar os dados com as predições de A e B. Este cálculo é simples para A, mas não para B, pois este último não pode fazer predições sem um valor para o parâmetro λ. Para contornar este problema, podemos expressar $P(D|B,I)$ em termos do parâmetro λ, usando a marginalização e a regra do produto:

$$P(D|B,I) = \int d\lambda P(D,\lambda|B,I) = \int d\lambda P(D|\lambda,B,I)P(\lambda|B,I) \tag{2.59}$$

O primeiro termo da integral $P(D|\lambda,B,I)$, onde o valor de λ é dado, é apenas uma função de verossimilhança comum. O segundo termo é a priori de B para λ. Para que os cálculos possam ser realizadas analiticamente, vamos admitir que o Sr. B pode apenas dizer que λ deva estar em algum intervalo entre λ_{min} e λ_{max}, originando assim uma distribuição a priori uniforme:

$$P(\lambda|B,I) = \frac{1}{\lambda_{max} - \lambda_{min}} \tag{2.60}$$

em que $\lambda_{min} \leq \lambda \leq \lambda_{max}$, sendo a priori igual a zero para valores de λ diferentes destes.

Vamos tambem admitir que há um valor λ_0 que mais se aproxima com as medidas, a probabilidade correspondente $P(D|\lambda_0,B,I)$ será o máximo da verossimilhança de B. De fato, quando o parâmetro ajustável está nas proximidades de seu valor ótimo, $\lambda_0 \pm \delta\lambda$, esperamos um *fitting* razoável dos dados, o que poderia ser representado por uma distribuição gaussiana:

$$P(D|\lambda,B,I) = P(D|\lambda_0,B,I)exp\left(-\frac{(\lambda-\lambda_0)^2}{2\delta\lambda^2}\right) \tag{2.61}$$

Substituindo os valores de $P(D|\lambda,B,I)$ e $P(\lambda|B,I)$ em $P(D|B,I)$ e levando em conta o fato de que a distribuição a priori não depende de λ e, portanto, pode ser retirada da integral, temos:

$$P(D|B,I) = \frac{1}{\lambda_{max} - \lambda_{min}} \int_{\lambda_{min}}^{\lambda_{max}} d\lambda P(D|\lambda,B,I) \tag{2.62}$$

Assim,

$$P(D|B,I) = \frac{P(D|\lambda_0,B,I) \times \delta\lambda\sqrt{2\pi}}{\lambda_{max} - \lambda_{min}} \tag{2.63}$$

Substituindo esta equação na expressão da razão das posteriores, obtemos:

$$\frac{P(A|D,I)}{P(B|D,I)} = \frac{P(A|I)}{P(B|I)} \times \frac{P(D|A,I)}{P(D|\lambda_0,B,I)} \times \frac{\lambda_{max} - \lambda_{min}}{\delta\lambda\sqrt{2\pi}} \tag{2.64}$$

O primeiro termo a direita reflete nossa inferência a priori para as teorias alternativas, a qual podemos tomar como sendo a unidade. O segundo termo é a medida de quão boa as melhores predições de cada modelo concordam com os dados; com a flexibilidade gerada por seu parâmetro ajustável, esta razão entre as verossimilhanças pode favorecer apenas a teoria B.

A análise da questão não se restringe, porém, só nisso - outro termo deve ser considerado. O intervalo a priori $\lambda_{max} - \lambda_{min}$ é geralmente muito maior que a incerteza $\pm\delta$ permitida pelos dados. Assim, o último termo da equação atua para penalizar B pelo parâmetro adicional, por esta razão ele é chamado de *fator de Occam*, inspirado na *Navalha de Occam*:"frusta fit per plura quod potest fiori per paciora", ou, "é em vão fazer com mais o que se pode ser feito com menos". Assim, na análise comparativa entre modelos, deve ser sempre balanceado fatores como o número de parâmetros que o modelo possui com seu grau de predição.

Na maioria dos casos, nossa preferência com relação a um modelo A ou outro B é dominada pela exatidão do ajuste com os dados, ou seja, a razão entre o máximo de suas verossimilhanças tende a ser o fator dominante. O fator de Occam pode tomar um papel principal, contudo, quando ambos os modelos dão resultados comparativamente bons com as medidas.

Outra característica importante surge quando consideramos o caso onde A também possui um parâmetro ajustável, μ por exemplo. Repetindo os mesmos procedimentos dos cálculos anteriores e fazendo as devidas simplicações, obtemos para esse caso:

$$\frac{P(A|D,I)}{P(B|D,I)} = \frac{P(A|I)}{P(B|I)} \times \frac{P(D|\mu_0,A,I)}{P(D|\lambda_0,B,I)} \times \frac{\delta\mu(\lambda_{max} - \lambda_{min})}{\delta\lambda(\mu_{max} - \mu_{min})} \tag{2.65}$$

Esta situação poderia representar, por exemplo, uma situação em que temos que escolher entre uma forma gaussiana ou lorentziana de um pico de sinal, mas cujo parâmetro seja desconhecido. A posição do máximo poderia ser fixada na origem por teoria e as amplitudes vinculadas pela normalização dos dados, os parâmetros μ e λ poderiam estar relacionados com a largura-a-meia-altura FWHM. Se atribuirmos igual peso para A e B antes da análise e um intervalo a priori similar para os parâmetros, então a equação anterior reduziria-se para:

$$\frac{P(A|D,I)}{P(B|D,I)} \approx \frac{P(D|\mu_0,A,I)}{P(D|\lambda_0,B,I)} \times \frac{\delta\mu}{\delta\lambda} \tag{2.66}$$

Para dados de boa qualidade, o fator dominante tenderia a ser a razão entre o melhor ajuste das verossimilhanças. Se ambos são comparáveis então o modelo com a maior barra-de-erro ($\delta\mu$ ou $\delta\lambda$) será favorecido. Isso parece razoável visto que uma barra-de-erro maior implica em que o parâmetro pode assumir mais valores que sejam consistentes com as hipóteses dadas. Exemplos de recentes aplicações em seleção de modelos podem ser encontras em análise de neutrinos observados a partir da supernova SN 1987A [18], comparações de modelos cosmológicos [19] e análise de dados coincidentes a partir de detectores de onde gravitacionais [20]

3

Princípio de Máxima Entropia

3.1 Medida de Entropia de Shannon

Toda distribuição de probabilidade tem associada consigo uma certa "incerteza". O conceito de "entropia"é introduzido aqui para fornecer uma medida quantitativa dessa incerteza. Shannon[12] sugeriu que tal medida de informação ou incerteza deveria obedecer à determinadas propriedades. Assim, sejam $p_1, p_2, .., p_n$ as probabilidades de ocorrência de n determinados eventos $A_1, A_2, ..., A_n$ de um dado experimento, dando origem à uma distribuição de probabilidades:

$$P = p(p_1, p_2, ..., p_n) \tag{3.1}$$

Obedecendo à condição de que $p_1 \geq 0, ..., p_n \geq 0$ e à condição de normalização:

$$\sum_{i=1}^{N} p_i = 1 \tag{3.2}$$

A medida de incerteza ou informação H deveria obedecer às seguintes propriedades:

1. H deveria ser uma função de $p_1, p_2, ..., p_n$ tal que ela possa ser escrita como:

$$H = H_n(P) = H_n(p_1, p_2, ..., p_n) \tag{3.3}$$

2. Deveria ser uma função contínua em $p_1, p_2, ..., p_n$

3. H deveria ser simétrica em relação aos seus argumentos, ou seja, quando os valores de saída ou eventos $A_1, A_2, ..., A_n$ são rearranjados entre si, H deve manter-se inalterada.

4. H não poderia variar caso seja acrescentado ao esquema de probabilidades um valor com probabilidade nula:
$$H_{n+1}(p_1, p_2, ..., p_n, 0) = H_n(p_1, p_2, ..., p_n) \tag{3.4}$$

5. H teria um mínimo e possivelmente um zero quando não há incerteza sobre o valor do evento, ou seja:
$$H_n(p_1, p_2, ..., p_n) = 0, \tag{3.5}$$
quando $p_i = 1, p_j = 0, i \neq j, i = 1, 2, ..., n$.

6. H deveria ter um valor máximo quando a incerteza é máxima, o que ocorre quando todos os eventos são igualmente prováveis, ou seja H é máximo quando:
$$p_i = ... = p_n = \frac{1}{n} \tag{3.6}$$

7. O valor máximo de H deveria aumentar quando n aumentasse, ou seja, a incerteza no processo é tanto maior quanto maior for o número de eventos possíveis.

8. Para duas distribuições de probabilidade independentes:
$$P = (p_1, p_2, ..., p_n), Q = (q_1, q_2, ..., q_m); \sum_{i=1}^{n} p_i = 1, \sum_{j=1}^{m} q_j = 1 \tag{3.7}$$

a incerteza do esquema adjunto PUQ deveria ser a soma de suas incertezas, ou seja:

$$H_{nm}(PUQ) = H_n(P) + H_m(Q) \tag{3.8}$$

onde se $A_1, A_2, ..., A_n$; $B_1, B_2, ..., B_n$ são os valores de saída de P e Q, então as saídas de PUQ são A_iB_j com probabilidades p_iq_j

Shannon sugeriu a seguinte medida:

$$H_n(p_1, p_2, ..., p_n) = -\sum_{i=1}^{n} p_i \ln p_i \tag{3.9}$$

É fácil notar que esta é uma função de $p_1, p_2, ..., p_n$, contínua e simétrica, se pudermos substituir $0ln(0)$ por zero. Ela não varia quando um evento de probabilidade nula é acrescentado ao esquema de probabilidade. Quando uma das probabilidades é igual a um, seu valor é zero, que é seu valor mínimo visto que $H \geq 0$ quando $0 \leq p_i \leq 1$. O valor máximo de H ocorre quando $p_i = 1/n$ e é igual a:

$$-\sum_{i=1}^{n} \frac{1}{n} ln\left(\frac{1}{n}\right) = ln(n) \tag{3.10}$$

e este valor, obviamente cresce quando n aumenta. A última propriedade pode ser demonstrada da seguinte forma:

$$H_{nm}(PUQ) = -\sum_{j=1}^{m}\sum_{i=1}^{n}(p_i q_j)ln(p_i q_j) = -\sum_{j=1}^{m} q_j[\sum_{i=1}^{n} p_i lnp_i] - \sum_{1=1}^{n} p_i[\sum_{j=1}^{m} q_j lnq_j] \tag{3.11}$$

Os termos em colchetes são respectivamente $H_n(P)$ e $H_m(Q)$; as somatórias $\sum_{j=1}^{m} q_j$ e $\sum_{1=1}^{n} p_i$, de modo que:

$$H_{nm}(PUQ) = H_n(P) + H_m(Q) \tag{3.12}$$

Portanto, a medida de Shannon satisfaz a todas as propriedades. Posteriormente, Kinchin [21] mostrou que qualquer medida que satisfaça à todas as propriedades descritas deve ter a forma:

$$-k\sum_{i=1}^{n} p_i lnp_i \tag{3.13}$$

onde k é uma constante positiva arbitrária. Outras medidas de incerteza podem ser obtidas somente modificando uma ou mais das regras ou mesmo retirando algumas e acrescentando outras.

3.2 O princípio de Máxima Entropia

Em 1957, E.T Jaynes enunciou o Princípio de Máxima Entropia, com o objetivo de derivar uma distribuição de probabilidade que descrevesse o microestado de um sistema, baseando-se em medições, ou seja, valores médios de funções, feitos em uma escala macroscópica. Jaynes buscou inspiração em argumentos da Teoria da Informação de Shannon para deduzir seu princípio e posteriormente aplicá-lo em problemas de mecânica estatística[13].

Em sua teoria, ele sugeriu uma reinterpretação da mecânica estatística que passaria de uma teoria física a um exemplo de inferência estatística. Neste sentido, a entropia de informação de Shannon teria um papel crucial mais básico que o conceito de energia. A partir desse princípio, estabeleceu algumas das propriedades fundamentais da inferência por máxima entropia e aplicou-a a mecânica estatística.

O problema básico do método de máxima entropia é o de encontrar a distribuição de probabilidades mais "honesta"possível dados os vínculos do problema, ou seja, a distribuição que melhor caracteriza o conjunto de dados sendo ao mesmo tempo coerente com os vínculos do sistema. Dada uma variável x que pode assumir valores discretos, não temos acesso às respectivas probabilidades individuais p_i, mas tão somente ao valor esperado de alguma função qualquer dessa variável x.

Segundo a teoria da informação, descrita em seus detalhes básicos na seção anterior, a distribuição que melhor representa o estado de informação disponível e que, ao mesmo tempo, seja a mais honesta ou mais "uniforme"é dada quando se maximiza a entropia de Shannon. Na seção seguinte, é feita uma demonstração de tal princípio, baseada na argumentação de Graham Wallis em 1962.

3.3 Uma Dedução para a Função de Máxima Entropia

Suponha que existam M possíveis valores de uma variável qualquer $\{x_i\}$: a questão proposta é a de atribuir valores de probabilidade para cada um dos valores de x, sabendo que para um dado x_i temos associado uma probabilidade $p\left(x_i|I\right) = p_i$, que corresponde a nossa informação testável, ou vínculo do problema . Vamos imaginar, para isso, um jogo em que cada valor de x é representado por uma caixa de igual tamanho.

Um time de macacos é escolhido para colocar em cada caixa um determinado número de moedas. Os hipotéticos macacos são usados para simbolizar que as escolhas das caixas estão sendo feitas aleatoramente. Depois de um grande número de moedas distribuídas, a fração em cada caixa representa a probabilidade associada a cada x_i. O resultado obtido pode não ser consistente com os vínculos de I, em cujo caso é descartado.

Após um grande número de tentativas, algumas distribuições são encontradas com mais frequência que as outras; a que ocorrer com mais frequência dentre todas e que satisfaça o vínculo imposto por I pode ser considerada como uma boa escolha para a distribuição de probabilidade $p\left(\{x_i\}\,|I\right)$. Vejamos como essa situação corresponde a encontrar uma distribuição com o maior valor de $-\sum_i p_i ln p_i$.

Depois que os macacos distribuíram todas as moedas nas caixas, encontraremos na primeira caixa n_1 moedas, na segunda n_2, etc. O número total de moedas N é então:

$$N = \sum_{i=1}^{M} n_i \tag{3.14}$$

onde assumiremos que $N >> M$. Dessa forma, temos:

$$p_i = n_i/N \tag{3.15}$$

Desde que cada moeda pode estar em qualquer caixa, há M^N maneiras de colocar as moedas nas caixas. Muitas dessas maneiras corresponderão a uma mesma distribuição de $\{n_i\}$. A frequência com a qual $\{p_i\}$ ocorrerá é dada por:

$$F\left(\{p_i\}\right) = \frac{N_{\{n_i\}}}{M^N} \tag{3.16}$$

onde $N_{\{n_i\}}$ é o número de maneiras de obter $\{n_i\}$.

Cálculo do numerador: tomando a caixa número 1, temos uma combinação de $C_{n_1}^N$ maneiras de escolher n_1 moedas dentre as N. Repetindo o mesmo raciocínio para as demais caixas, chegamos em:

$$C_{n_1}^N . C_{n_2}^{N-n_1} . C_{n_3}^{N-n_1-n_2} .... C_{n_M}^{n_M} = \frac{N!}{n_1! n_2! ... n_M!} \tag{3.17}$$

Portanto, a frequência com que é obtido $\{p_i\}$ é dada por:

$$F = \frac{N!}{n_1! n_2! ... n_M!} \frac{1}{M^N} \tag{3.18}$$

Tomando o lnF :

$$lnF = ln(N!) - Nln(M) - \sum_{i=1}^{M} ln(n_i) \tag{3.19}$$

O lado direito pode ser simplificado, usando a aproximação de stirling: $ln(n!) \approx lnn - n$, quando $n \to \infty$.

Desta forma:

$$lnF = -NlnM + NlnN - \sum_{i=1}^{M} ln(n_i) \tag{3.20}$$

Usando a relação $p_i = n_i/N$ e o fato de que $\sum_i p_i = 1$, temos que:

$$lnF = -NlnM - N\sum_i p_i lnp_i \tag{3.21}$$

A probabilidade, $p(\{x_i\}|I)$, que buscamos é aquela que tem a maior frequência F. Assim, como N e M são constantes, obter o maior valor de F ou lnF significa maximizar a função:

$$S = -\sum_i p_i lnp_i \tag{3.22}$$

Como vimos, buscar uma distribuição de probabilidade baseada apenas em informação disponível é equivalente a maximizar a entropia de Shannon.

3.3.1 Exemplo Elementar da Aplicação do Método de Máxima Entropia

Para verificar um paralelo entre o que nossa intuição pode dar como resposta a um determinado problema e o que o método de máxima entropia fornece como resultado, abaixo está citado um exemplo bastante trivial conhecido como o *problema do canguru*. Dado um conjunto de cangurus sabe-se que 1/3 deles têm olhos azuis e 1/3 são canhotos. Com base nestas informações, nossa questão é que proporção de todos os cangurus são canhotos e têm os olhos azuis..

Podemos sistematizar o problema da seguinte forma: vamos escrever cada uma das possibilidades distintas da seguinte forma:

- p_1 para cangurus de olhos azuis e canhotos
- p_2 olho-azul e destro
- p_3 para cangurus canhotos e que não tenham olhos azuis
- p_4 cangurus destros e que não tenham olhos azuis

Note que estas probabilidades não são independentes. Há vínculos impostos pelas informações que temos do problema. Assim, podemos escrever :

$$p_1 + p_2 = \frac{1}{3} \tag{3.23}$$

e

$$p_1 + p_3 = \frac{1}{3} \tag{3.24}$$

Além dessas duas, temos a condição de normalização :

$$p_1 + p_2 + p_3 + p_4 = 1 \tag{3.25}$$

Manipulando as equações, chegamos a conclusão que todas as soluções onde $0 \leq x \leq 1/3$ satisfazem o problema, onde $x = p_1$ Então, qual seria a solução mais adequada com as informações que tempos em mãos ? Se uma escolha tivesse de ser feita dentre todas as soluções possíveis, nosso senso comum escolheria aquela que fosse a mais independente possível, ou seja, $x = 1/9$. Qualquer valor diferente desse indicaria haver alguma correlação entre a cor dos olhos e o fato dos cangurus serem canhotos ou não; no entanto, não possuimos tal informação. Skilling [22] mostrou que as únicas funções que resultam em $x = 1/9$ são aquelas relacionadas monotonicamente com a função de entropia:

$$S = -\sum_{i=1}^{4} p_i ln p_i = -xlnx - 2\left(\frac{1}{3} - x\right) ln\left(\frac{1}{3} - x\right) - \left(\frac{1}{3} + x\right) ln\left(\frac{1}{3} + x\right) \tag{3.26}$$

Para ilustrar esse ponto, na tabela abaixo mostramos os resultados de maximização do valor de x utilizando outras "três funções de entropia":

A partir da análise da tabela, observamos que dentre as quatro funções variacionais propostas, a única que não apresenta nenhum valor de x que leve à alguma correlação é a função de entropia. As demais apresentam correlações envolvidas, ou seja, as correlações positivas indicariam que a cor dos olhos teriam uma relação direta com o fato dos canhotos serem canhotos, e a correlação negativa indicaria o oposto; nenhuma dessas informaçãoes está contida no nosso problema, portanto, pode-se perceber aqui que a função de máxima entropia adequada é a proposta por Shannon.

Tabela 3.1: Soluções para o problema dos cangurus quando se maximiza quatro funções diferentes, sujeitas aos vínculos do problema.

Função Variacional	**Valor máximo de x**	**Correlação Implicada**
$-\sum_i p_i ln p_i$	0.1111	nenhuma
$-\sum_i p_i^2$	0.0833	negativa
$\sum_i ln p_i$	0.1301	positiva
$\sum_i \sqrt{p_i}$	0.1218	positiva

3.4 O Formalismo do Método de Maximização de Entropia

Dados alguns valores médios, há infinitas distribuições de probabilidade compatíveis. De acordo com a MaxEnt, deve-se selecionar dentre todas as distribuições possíveis aquela que maximize a entropia de Shannon, sendo ao mesmo tempo compatível com os vínculos impostos pelos valores médios e condições de normalização.

Assim, consideremos uma variável randômica X que assuma os valores $x_1, x_2, ..., x_n$ com as respectivas probabilidades associadas $p_1, p_2, ..., p_n$. Com tais valores de probabilidade, podemos encontrar valores médios de funções de X, $g_1(X), g_2(X), ..., g_m(X)$.

Desta forma, podemos escrever as equações de vículo como sendo:

$$\sum_{i=1}^{n} p_i g_r(x_i) = a_r \tag{3.27}$$

onde r=1,2,...m

Além das equações de vínculo acima, há a condição de normalização:

$$\sum_{i=1}^{n} p_i = 1 \tag{3.28}$$

Como pode ser visto, temos $m+1$ equações para a determinação de $p_1, p_2, ..., p_n$. Em geral, $m+1 < n$ de modo que há um infinito número de soluções para o sistema. O método de Máxima Entropia sugere que dentre todas as soluções, devemos escolher a que maximiza a entropia.

Para maximizar a entropia, usaremos o método de Lagrange[23]. Assim, dados os vínculos e a condição de normalização, construímos a Lagrangiana:

$$L \equiv -\sum_{i=1}^{n} p_i ln p_i - (\lambda_0 - 1)\left(\sum_{i=1}^{n} p_i - 1\right) - \sum_{r=1}^{m} \lambda_r \left(\sum_{i=1}^{r} p_i g_{ri} - a_r\right) \tag{3.29}$$

onde $\lambda_0, \lambda_1, ..., \lambda_m$ são os multiplicadores de Lagrange; note que $\lambda_0 - 1$ é usado no lugar de λ_0 por questão de conveniência.

Tomando a derivada de L com relação a p_i e igualando a zero, temos:

$$\frac{\partial L}{\partial p_i} = 0 \tag{3.30}$$

$$-ln p_i - \lambda_0 - \sum_{r=1}^{m} \lambda_r g_{ri} = 0 \tag{3.31}$$

E, portanto:

$$p_i = exp\left(-\lambda_0 - \lambda_1 g_{1i} - \lambda_2 g_{2i} - .... - \lambda_m g_{mi}\right) \tag{3.32}$$

Para determinarmos os valores dos multiplicadores de Lagrange, substituimos (3.32) em (3.27) e (3.28), tal que:

$$\sum_{i=1}^{n} exp\left(-\lambda_0 - \sum_{j=1}^{m} \lambda_j g_{ji}\right) = 1 \tag{3.33}$$

e

$$\sum_{i=1}^{n} g_{ri} exp\left(-\lambda_0 - \sum_{j=1}^{m} \lambda_j g_{ji}\right) = a_r \tag{3.34}$$

sendo que $r = 1, 2, ..., m$

de tal forma que:

$$exp\left(\lambda_0\right) = \sum_{i=1}^{n} exp\left(-\sum_{j=1}^{m} \lambda_j g_{ji}\right) \tag{3.35}$$

e

$$a_r = \frac{\sum_{i=1}^{n} g_{ri} exp\left(-\sum_{j=1}^{m} \lambda_j g_{ji}\right)}{\sum_{i=1}^{n} exp\left(-\sum_{j=1}^{m} \lambda_j g_{ji}\right)} \tag{3.36}$$

A equação (3.35) dá o valor de λ_0 como uma função de $\lambda_1, \lambda_2, ..., \lambda_m$, enquanto as equações (3.36) calculam $a_1, a_2, ..., a_m$ como funções de $\lambda_1, \lambda_2, ..., \lambda_m$.

O formalismo acima descreve a dedução de uma distribuição de probabilidade genérica a partir de vínculos gerais. Na prática, esses vínculos dependerão do problema a ser considerado.

3.5 Verificação de que o Método de Lagrange origina um Máximo Global de Entropia

Para a verificação completa de que o extremo obtido pelo método de lagrange é, de fato, um valor de máximo global, procedemos como segue. Seja:

$$S = -\sum_{i=1}^{n} p(x_i) ln p(x_i) \tag{3.37}$$

Sujeita aos vínculos:

$$\sum_{i=1}^{n} p(x_i) = 1 \tag{3.38}$$

$$\sum_{i=1}^{n} p(x_i) g_r(x_i) = \bar{g}_r(x_i) \tag{3.39}$$

com $r = 1, 2, ..., m$.

A solução é dada por:

$$p(x_i) = exp[-\lambda_0 - \lambda_1 g_1(x_1) - ... - \lambda_m g_m(x_i)] \tag{3.40}$$

Seja F a entropia para qualquer outra distribuição de probabilidade que satisfaça aos mesmos vínculos dados anteriormente, de forma que:

$$F = -\sum_{i=1}^{n} f(x_i) ln f(x_i) \tag{3.41}$$

$$\sum_{i=1}^{n} f(x_i) = 1 \tag{3.42}$$

$$\sum_{i=1}^{n} f(x_i) g_r(x_i) = \bar{g}_r(x_i) \tag{3.43}$$

Desta forma, queremos calcular a quantidade $S_{max} - F$, e verificar o sinal desta função. Assim:

$$S_{max} - F = -\sum_{i=1}^{n} p(x_i) lnp(x_i) + \sum_{i=1}^{n} f(x_i) lnf(x_i) \tag{3.44}$$

$$= \sum_{i=1}^{n} [f(x_i) - p(x_i)] lnp(x_i) + \sum_{i=1}^{n} f(x_i) \left[lnf(x_i) - lnp(x_i) \right] \tag{3.45}$$

Tomando o valor de $p(x_i)$ dado pela equação 3.40, e substituindo na expressão, temos:

$$S_{max} - F = \sum_{i=1}^{n} [f(x_i) - p(x_i)][-\lambda_0 - \lambda_1 g_1(x_1) - ... - \lambda_m g_m(x_i)] + \sum_{i=1}^{n} f(x_i) ln \frac{f(x_i)}{p(x_i)} \tag{3.46}$$

Cada termo do primeiro somatório se anula pela condição de normalização e, assim, resta analisar a última somatória. Para tanto, vamos usar uma propriedade das funções convexas ($xlnx$ é um exemplo de função convexa), chamado de *desigualdade de Jensen* :

$$E[\phi(x)] \geq \phi[E(x)] \tag{3.47}$$

Assim, dada duas distribuições de probabilidade quaisquer $P = (p_1, p_2, ..., p_n)$ e $Q = (q_1, q_2, .., q_n)$ e tomando $\phi(x) = x \ln x$ e $x = p_i/q_i$ com probabilidades de ocorrência dada por $q_i (i = 1, 2, ..., n)$, então a desigualdade de Jensen é expressa da seguinte forma:

$$\sum_{i=1}^{n} \frac{p_i}{q_i} ln \frac{p_i}{q_i} q_i \geq \sum_{i=1}^{n} q_i \frac{p_i}{q_i} ln \left[\sum_{i=1}^{n} q_i \frac{p_i}{q_i} \right] \tag{3.48}$$

Fazendo as devidas simplicações e usando o fato de que $\sum_{i=1}^{n} p_i = 1$, temos, finalmente:

$$\sum_{i=1}^{n} p_i ln \frac{p_i}{q_i} \geq 0 \tag{3.49}$$

Usando este resultado no nosso problema original, chegamos à conclusão de que $S_{max} - F \geq 0$, ou seja, S_{max} é um máximo global e $F = S_{max}$ quando $p(x_i) = f(x_i)$ para todo i.

3.6 Entropia como Ferramenta de Indução

Apresentaremos a seguir um formalismo diferente do proposto por Shannon para a derivação de uma forma funcional da entropia [24]. Esta foi motivada por algumas questões que referiam-se ao fato de a entropia de Shannon tratar apenas dos casos discretos e, além disso, buscava-se uma forma que não dependesse de interpretações do tipo "quantidade de incerteza"ou "informação". Neste sentido, é importante a contribuição de Shore e Johnson[14], que buscaram um metódo axiomático que focalizasse sua atenção no método de inferência em si, e não na medida de informação. Por esse formalismo, a entropia perde a interpretação de quantidade de incerteza ou informação, tornando-se meramente uma ferramenta de indução.

Para tanto considere uma variável x em um espaço X, em que x pode ser discreto ou contínuo, em uma ou mais dimensões. A incerteza sobre o valor de x está relacionada a uma distribuição de probabilidades $q(x)$. A questão é a de atualizar a distribuição a priori $q(x)$ em uma distribuição posterior $p(x)$ quando novas informações em forma de vínculos tornam-se disponíveis.

Para selecionar uma distribuição posterior dentre uma família de distribuições compatíveis com os vínculos torna-se desejável ordenar tais distribuições de uma maneira transitiva, ou seja, se dada distribuição p_1 é preferível a uma distribuição p_2, e p_2 preferível a uma distribuição p_3, então p_1 é preferível a p_3. Tal ordenamento é realizado atribuindo a cada $p(x)$ um número real $S(p)$ de tal forma que, se p_1 é preferível a p_2 então $S(p_1) > p_2$. A distribuição selecionada no final será aquela que maximize o funcional $S(p)$, que será chamado de entropia de p. A forma funcional de $S(p)$ deve respeitar alguns axiomas.

Axioma 1: Localidade: *Informação local tem efeitos locais.* Supondo que a informação a ser processada refira-se a apenas um sub-domínio D de X nada dizendo sobre os valores de x fora de D. A função de entropia deve ser tal que a probabilidade de qualquer x fora de D, $p(x|X \notin D)$ não seja modificada, visto que não há informação disponível para isso. A consequência do axioma 1 é que domínios não-sobrepostos de x contribuem aditivamente com a entropia:

$$S(p) = \int F(p(x), x)dx \tag{3.50}$$

Axioma 2: Invariância de Coordenada: *O sistema de coordenadas não carrega nenhuma informação.* Os pontos x podem ser escritos usando qualquer sistema de coordenadas, sendo que o

ordenamento das distribuiçoes de probabilidade não devam depender do sistema de coordenadas usado. A consequência do axioma 2 é que $S(p)$ pode ser escrito em termos de invariantes por mudança de coordenadas, tais como $dxm(x)$ e $p(x)/m(x)$:

$$S(p) = \int m(x)\phi(x)\frac{p(x)}{m(x)}dx \tag{3.51}$$

A densidade $m(x)$ e a função $\phi(x)$ ainda estão indeterminadas. No entanto, usando um caso especial do axioma 1 pode-se especificar a natureza de $m(x)$.

Axioma 1 (Caso especial): Quando não há nenhuma nova informação não há razão para uma atualização da distribuição a priori que deve coincidir com a distribuição posterior. Assim, como consequência, $m(x)$ pode ser tomado, a menos de uma constante de normalização, como a distribuição a priori.

Axioma 3: Independência de Subsistemas: *Quando um sistema é composto por subsistemas em que há razão para serem considerados como independentes não importa se o procedimento de inferência trate-os separadamente ou conjuntamente.*

Considere um sistema composto por dois subsistemas, $x = (x_1, x_2) \in X = X_1 \times X_2$. Assuma que as evidências a priori indiquem que os sistemas são independentes, de forma que se as distribuições a priori dos subsistemas são $m_1(x_1)$ e $m_2(x_2)$, então a distribuição a priori para o sistema todo é $m_1(x_1)m_2(x_2)$. Depois, suponha que nova informação é adquirida tal que $m_1(x_1)$ é atualizada para $p_1(x_1)$ e $m_2(x_2)$ para $p_2(x_2)$.

Baseando-se apenas nessas novas informações, não há razões para revisar-se a visão prévia de independência entre os subsistemas, de modo que a distribuição a priori para o sistema todo $m_1(x_1)m_2(x_2)$ pode ser atualizada para a distribuição posterior $p_1(x_1)p_2(x_2)$. Isto é uma propriedade típica de funções do tipo $logx$. Como consequência do axioma 3 e, resultado final, temos que as distribuiçõees de probabilidade $p(x)$ deveriam ser ordenadas relativamente a distribuição a priori $m(x)$ de acordo com sua *entropia relativa*:

$$S(p|m) = -\int p(x)ln\frac{p(x)}{m(x)}dx \tag{3.52}$$

No caso de uma variável discreta, se atribuirmos iguais probabilidades a priori, $m_i = 1$, a entropia é dada por:

$$S(p) = -\sum_i p_i ln p_i \tag{3.53}$$

Ou seja, a entropia de Shannon é um caso particular da Eq. 3.52. É necessário enfatizar que mesmo quando a entropia de Shannon é usada, a medida a priori $m_i = 1$ está sendo levada em conta implicitamente.

3.7 Consistência entre os Métodos de Bayes e o de Máxima Entropia

Para reforçar o conceito de atualização das probabilidades, vamos mudar ligeiramente a notação, e escrever índices "velho"e "novo"em cada probabilidade para deixar explícito que houve um processo de atualização de nossa inferência sobre detrminado parâmetro[25]. Assim, a regra de Bayes pode ser reescrita como:

$$P_{novo}(\theta) = P(\theta|X) = P_{velho}(\theta)\frac{P(X|\theta)}{P_{velho}(X)} \tag{3.54}$$

Onde denotamos por X os valores dos dados x que foram observados, e θ algum parâmetro de interesse. Com esta expressão reescrita, vamos agora verificar a consistência entre os dois métodos. Vimos que a aplicação do método de Máxima Entropia atualiza em uma distribução posterior uma distribuição a priori, dada a informação disponível. Assim, a distribuição posterior selecionada $P_{novo}(x, \theta)$ será aquela que maximize:

$$S[P, P_{velho}] = -\int dx d\theta P(x, \theta) ln \frac{P(x, \theta)}{P_{velho}(x, \theta)} \tag{3.55}$$

A informação a ser processada (os valores observados X) devem ser expressados na forma de vínculos. Claramente, a família de distribuições de probabilidade posteriores que reflete o fato de que x é agora conhecido e toma os valores de X é tal que:

$$P(x) = \int d\theta P(x, \theta) = \delta(x - X) \tag{3.56}$$

Impomos também uma condição de normalização:

$$\int dxd\theta P(x,\theta) = 1 \tag{3.57}$$

Assim, procedemos escrevendo a Lagrangiana:

$$L = S + \int dx\lambda(x)[\int d\theta P(x,\theta) - \delta(x-X)] + \alpha[\int dxd\theta P(x,\theta) - 1] \tag{3.58}$$

onde $\lambda(x)$ e α são os multiplicadores de Lagrange. Devemos agora derivar a expressão anterior com relação a $P(x,\theta)$ e igualar a zero. Deste modo, calculando a derivada de cada um dos três itens da lagrangiana, temos:

$$\frac{\partial S}{\partial P(x,\theta)} = -\int dxd\theta \frac{\partial}{\partial P(x,\theta)}[P(x,\theta) ln \frac{P(x,\theta)}{P_{velho}(x,\theta)}] \tag{3.59}$$

Portanto:

$$\frac{\partial S}{\partial P(x,\theta)} = \int dxd\theta [ln \frac{P(x,\theta)}{P_{velho}(x,\theta)} + 1] \tag{3.60}$$

Derivada do segundo termo:

$$\frac{\partial}{\partial P(x,\theta)} \int dx\lambda(x)[\int d\theta P(x,\theta) - \delta(x-X)] = \int dxd\theta\lambda(x) \tag{3.61}$$

E, finalmente, a derivada do terceiro termo:

$$\frac{\partial}{\partial P(x,\theta)} \alpha[\int dxd\theta P(x,\theta) - 1] = \alpha \int dxd\theta \tag{3.62}$$

Assim, somando as derivadas e igualando a zero:

$$\int dxd\theta [-ln \frac{P(x,\theta)}{P_{velho}(x,\theta)} + \lambda(x) + \alpha - 1] = 0 \tag{3.63}$$

Cuja solução geral é dada quando o valor do integrando é identicamente nulo, portanto:

$$-ln \frac{P(x,\theta)}{P_{velho}(x,\theta)} + \lambda(x) + \alpha - 1 = 0 \tag{3.64}$$

E, resolvendo para $P(x,\theta)$:

$$P(x,\theta) = P_{velho}(x,\theta) e^{\lambda(x)+\alpha-1} \tag{3.65}$$

Note que podemos escrever:

$$Z = e^{\alpha-1} = \int dx d\theta P_{velho}(x,\theta) e^{\lambda(x)} \tag{3.66}$$

E assim, reescrevemos a Eq. (3.65) como:

$$P_{novo}(x,\theta) = P_{velho}\frac{e^{\lambda(x)}}{Z} \tag{3.67}$$

Os multiplicadores são obtidos a partir da Eq. (3.56),

$$\int d\theta P_{velho}(x,\theta)\frac{e^{\lambda(x)}}{Z} = P_{velho}(x,\theta)\frac{e^{\lambda(x)}}{Z} = \delta(x-X) \tag{3.68}$$

Então, substituindo $e^{\lambda(x)}$ de volta em (3.67):

$$P_{novo}(x,\theta) = \frac{P_{velho}(x,\theta)\delta(x-X)}{P_{velho}(x)} \tag{3.69}$$

Relembrando a regra do produto, temos:

$$P_{velho}(x,\theta) = P_{velho}(x)P_{velho}(\theta|x) \tag{3.70}$$

Substituindo o termo $P_{velho}(x,\theta)/P_{velho}(x)$ por $P_{velho}(\theta|x)$, temos:

$$P_{novo}(x,\theta) = \delta(x-X)P_{velho}(\theta|x) \tag{3.71}$$

A nova distribuição marginal para θ é:

$$P_{novo}(\theta) = \int dx P_{novo}(x,\theta) = \int dx \delta(x-X)P_{velho}(\theta|x) = P_{velho}(\theta|X) \tag{3.72}$$

que é a regra de Bayes. Assim, a atualização via Bayes é um caso especial de atualização via Máxima Entropia.

Resumidamente: a probabilidade a priori $P_{velho}(x,\theta) = P_{velho}(x)P_{velho}(\theta|x)$ é atualizada para a posterior $P_{novo}(x,\theta) = P_{novo}(x)P_{novo}(\theta|x)$ onde $P_{novo}(x) = \delta(x-X)$ é fixado pelos dados observados enquanto $P_{novo}(\theta|x) = P_{velho(\theta|x)}$ mantem-se inalterado. Note que isto está em acordo com a filosofia que orienta o método de Máxima Entropia, *atualiza-se apenas aqueles aspectos de nossas crenças para as quais novas evidências corretivas tenham sido fornecidas* . A generalização para situações

onde há alguma incerteza sobre os valores dos dados atuais é simples, sendo necessário substituir a função δ de $P(x)$ por alguma distribuição conhecida.

3.8 Aplicação do Método em Mecânica Estatística

Os primeiros resultados do método de máxima entropia foram em mecânica estatística, realizados também por Jaynes, em seu artigo original. Por essa razão, mostraremos aqui a derivação de algumas das principais distribuições,Maxwell-Boltzmann,Bose-Einstein e Fermi-Dirac, com base na aplicação do princípio de máxima entropia.

3.8.1 Distribuição de Maxwell-Boltzmann

Nessa distribuição, temos como informação disponível o valor médio da energia do sistema. Baseado unicamente nessa informação, aplicaremos o princípio de MaxEnt para derivar a distribuição. Sendo $p_1, p_2, ..., p_n$ as probabilidades que uma partícula tenha energias $\epsilon_1, \epsilon_2, ..., \epsilon_n$, sua energia média $\hat{\epsilon}$ é dada por:

$$p_1\epsilon_1 + p_2\epsilon_2 + ... + p_n\epsilon_n = \hat{\epsilon} \tag{3.73}$$

Temos, além disso, a condição de normalização:

$$p_1 + p_2 + ... + p_n = 1 \tag{3.74}$$

onde $p_i \geq 0$ para todo i

Para aplicar o princípio de MaxEnt, escrevemos a Lagrangiana desse sistema, como sendo:

$$L \equiv -\sum_{i=1}^{n} p_i lnp_i - (\lambda - 1)\left(\sum_{i=1}^{n} p_i - 1\right) - \mu\left(\sum_{i=1}^{n} p_i\epsilon_i - \hat{\epsilon}\right) \tag{3.75}$$

Tomando as derivadas de L com respeito a p_i iguais a zero, obtemos:

$$-lnp_i - \lambda - \mu\epsilon_i = 0 \tag{3.76}$$

ou

$$p_i = e^{-\lambda - \mu\epsilon_i} \tag{3.77}$$

Substituindo a equação acima na condição de normalização, $\sum_i p_i = 1$, temos que :

$$e^{-\lambda} = \left(\sum_{i=1}^{n} e^{-\mu\epsilon_i} \right)^{-1} \tag{3.78}$$

Assim, finalmente obtemos:

$$p_i = \frac{e^{-\mu\epsilon_i}}{\sum_{i=1}^{n} e^{-\mu\epsilon_i}} \tag{3.79}$$

Esta é a distribuição de Maxwell-Boltzmann da mecânica estatística. Para obtê-la, aplicando MaxEnt, usamos apenas o vínculo imposto pela energia média do sistema de partículas. Vejamos agora as demais distribuições.

3.8.2 Distribuição de Bose-Einstein

Nesse caso, vamos adicionar um vínculo ao problema. Além do conhecimento da energia média do sistema, vamos supor que conhecemos o número médio de partículas $\bar{N}$, (o número total de partículas pode variar de zero a infinito). Assim sendo, denotaremos por p_{ij} a probabilidade de encontrar j partículas no i-ésimo estado, de um total de n estados. Assim:

$$\sum_{j=0}^{\infty} p_{ij} = 1 \tag{3.80}$$

onde i=1, 2, .., n.

O número médio de partículas e a energia esperada do sistema são dadas respectivamente por:

$$\sum_{i=1}^{n} \sum_{j=0}^{\infty} j p_{ij} = \bar{N} \tag{3.81}$$

e

$$\sum_{i=1}^{n} \epsilon_i \sum_{j=0}^{\infty} j p_{ij} = \hat{\epsilon} \tag{3.82}$$

A entropia total do sistema é dada somando-se as entropias individuais sobre todos os estados:

$$-\sum_{i=1}^{n} \sum_{j=0}^{\infty} p_{ij} ln p_{ij} \tag{3.83}$$

O passo seguinte é construir a lagrangiana, usando os vínculos e a condição de normalização:

$$L \equiv -\sum_{i=1}^{n}\sum_{j=0}^{\infty} p_{ij} ln p_{ij} - \lambda\left(\sum_{i=1}^{n}\epsilon_i \sum_{j=0}^{\infty} j p_{ij} - \bar{N}\right) - \mu\left(\sum_{i=1}^{n}\epsilon_i \sum_{j=0}^{\infty} j p_{ij} - \hat{\epsilon}\right) - \sum_{i=1}^{n}\nu_i\left(\sum_{j=0}^{\infty} p_{ij} - 1\right) \tag{3.84}$$

Com a lagrangiana em mãos, tomamos as derivadas de L com relação aos p_{ij} e rearranjamos os termos, tal que obtemos :

$$p_{ij} = \left(1 - e^{-\lambda-\mu\epsilon_i}\right) exp\left[-j\left(\lambda + \mu\epsilon_i\right)\right] \tag{3.85}$$

O número médio de partículas no i-ésimo estado é dado, então, por:

$$\bar{N}_i = \sum_{j=0}^{\infty} j p_{ij} = \left(1 - e^{-\lambda-\mu\epsilon_i}\right)\sum_{j=0}^{\infty} j e^{-j(\lambda+\mu\epsilon_i)} = \frac{1}{e^{\lambda+\mu\epsilon_i} - 1} \tag{3.86}$$

A distribuição $\left(\bar{N}_1, \bar{N}_2, \bar{N}_3, ...\right)$ do número de partículas nos n estados é chamada de Distribuição de Bose-Einstein.

3.8.3 Distribuição de Fermi-Dirac

Nesta distribuição, um vínculo é adicionado ao problema. Ao contrário da distribuição de Bose-Einstein, aqui cada estado pode ser ocupado por apenas uma ou nenhuma partícula. Portanto, nesse caso, temos que maximizar:

$$-\sum_{i=1}^{n}\sum_{j=0}^{1} p_{ij} ln p_{ij} \tag{3.87}$$

sujeita à condição de normalização:

$$\sum_{j=0}^{1} p_{ij} = 1 \tag{3.88}$$

E aos vínculos impostos pelos valores médio de partícula e energia, respectivamente :

$$\sum_{i=1}^{n}\sum_{j=0}^{1} j p_{ij} = \bar{N} \tag{3.89}$$

e

$$\sum_{i=1}^{n} \epsilon_i \sum_{j=0}^{1} j p_{ij} = \hat{\epsilon} \tag{3.90}$$

Aplicando o método de Lagrange, obtemos a seguinte probabilidade:

$$p_i = c_i exp\left[-j\left(\lambda + \mu\epsilon_i\right)\right] \tag{3.91}$$

Usando a condição de normalização, podemos determinar assim o número médio de partículas no i-ésimo estado:

$$\bar{N}_i = \sum_{j=0}^{1} j p_{ij} = p_{i1} = c_i e^{-\lambda-\mu\epsilon_i} = \frac{e^{-\lambda-\mu\epsilon_i}}{e^{-\lambda-\mu\epsilon_i}+1} = \frac{1}{e^{\lambda+\mu\epsilon_i}+1} \tag{3.92}$$

Essa é a chamada distribuição de Fermi-Dirac.

3.9 Algoritmo que produz numericamente distribuições de máxima entropia

Nesta seção descreveremos um método computacional [26] com o intuito de gerar distribuições de máxima entropia, dados os valores das funções e dos vínculos. A formulação geral do problema de máxima entropia é dada como segue:

- Temos que maximizar:

$$S = -\int p(x) ln\left[\frac{p(x)}{m(x)}\right] dx \tag{3.93}$$

- Sujeito as vínculos:

$$E[\phi_n(x)] = \int \phi_n(x) p(x) dx = \mu_n \tag{3.94}$$

  onde $\mu_0 = 1$, $\phi_0(x) = 1$ e $\phi_n(x)$, $n = 0, ..., N$ são N funções desconhecidas, e μ_n, $n = 0, ..., N$ são os valores esperados ou vínculos.

- A solução geral do problema é dada por:

$$p(x) = m(x)exp\left[-\sum_{n=0}^{N}\xi_n\phi_n(x)\right] \tag{3.95}$$

- Os $(N+1)$ multiplicadores de Lagrange representados aqui por $\left[\vec{\xi} = \xi_0, ..., \xi_n\right]$ são obtidos resolvendo o conjunto de $(N+1)$ equações não-lineares:

$$G_n(\vec{\xi}) = \int m(x)\phi_n(x)exp\left[-\sum_{n=0}^{N}\xi_n\phi_n(x)\right]dx = \mu_n, n = 0, ..., N \tag{3.96}$$

Estas equações podem ser resolvidas pelo método de Newton, que consiste em expandir $G_n(\vec{\xi})$ em séries de Taylor em torno de um valor inicial para $\vec{\xi}$, desprezando os termos de ordem quadrática e superior, e resolvendo o sistema linear resultante iterativamente. O desenvolvimento em primeira ordem da série de Taylor de $G_n(\vec{\xi})$ resulta em:

$$G_n(\vec{\xi}) \cong G_n(\vec{\xi_0}) + (\vec{\xi} - \vec{\xi_0})^t[\mathbf{Grad}G_n(\vec{\xi})]_{(\vec{\xi}=\vec{\xi_0})} = \mu_n \tag{3.97}$$

Definindo os vetores $\vec{\delta}$ e $\vec{v}$ por:

$$\vec{\delta} = \vec{\xi} - \vec{\xi_0} \tag{3.98}$$

$$\vec{v} = [\mu_0 - G_0(\vec{\xi_0}), ..., \mu_N - G_N(\vec{\xi_0})]^t \tag{3.99}$$

E a matriz $\vec{G}$ por:

$$\vec{G} = (g_{nk}) = \left(\frac{\partial G_n(\vec{\xi})}{\partial \xi_k}\right)_{\vec{\xi}=\vec{\xi_0}} \tag{3.100}$$

Desse modo, a Eq. 3.97 torna-se:

$$\vec{G}.\vec{\delta} = \vec{v} \tag{3.101}$$

Note que a matriz $\vec{G}$ é simétrica e, assim:

$$g_{nk} = g_{kn} = -\int m(x)\phi_n(x)\phi_k(x)exp[-\sum_{n=0}^{N}\xi_n\phi_n(x)]dx \tag{3.102}$$

Portanto, para a determinação dos g_{nk} é necessário o cálculo de $N(N-1)/2$ integrais, além dessas, há $N+1$ integrais para $\vec{G}$; as integrais sendo calculadas por algum método numérico apropriado, neste caso, o método de quadratura de Gauss-Legendre.

3.9.1 Distribuição Exponencial

Iniciaremos a dedução de algumas distribuições clássicas de probabilidade com a distribuição exponencial. A seguir fazemos sua dedução analítica, posteriormente mostrando o resultado do programa. Para caracterizá-la como uma distribuição de máxima entropia, temos que considerar os seguintes vínculos:

$$\int_0^\infty p(x)d(x) = 1 \tag{3.103}$$

$$\int_0^\infty xp(x) = \mu_1 = E(x) \tag{3.104}$$

A forma geral da distribuição de máxima entropia gerada a partir destes vínculos e valores médios é:

$$p(x) = m(x)e^{-\xi_0 - \xi_1 x} \tag{3.105}$$

Consideraremos $m(x) = 1$, restando resolver então um sistema de duas equações:

$$\int_0^\infty e^{-\xi_0 - \xi_1 x}dx = 1 \tag{3.106}$$

$$\int_0^\infty xe^{-\xi_0 - \xi_1 x}dx = \mu_1 \tag{3.107}$$

Calculando as integrais e resolvendo o sistema, temos:

$$e^{-\xi_0} = \frac{1}{\mu_1} \rightarrow \xi_0 = ln\mu_1 \tag{3.108}$$

$$\xi_1 = \frac{1}{\mu_1} \tag{3.109}$$

Deste modo, a distribuição exponencial é dada por:

$$p(x) = e^{-ln\mu_1 - \frac{x}{\mu_1}} \tag{3.110}$$

onde μ_1 é o valor médio de x.

Na tabela 3.2 mostramos alguns casos numéricos, em que variamos o valor médio de x. Na Fig. 3.1, apresentamos as distribuições resultantes.

Tabela 3.2: Distribuição Exponencial.

μ_1	λ_0	λ_1
1.0	0.0024	0.9973
2.0	0.7255	0.4801
3.5	1.2578	0.2840

3.9.2 Distribuição Normal

Quando é o momento de segunda ordem é prescrito com relação à origem $E[x^2] = \mu_2$, e denotando-o por σ^2, temos:

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{3.111}$$

$$\int_{-\infty}^{\infty} x^2 p(x)dx = \sigma^2 \tag{3.112}$$

A forma geral da função de entropia é dada por:

$$p(x) = e^{-\xi_0 - \xi_1 x^2} \tag{3.113}$$

Procedemos da mesma forma como a feita no caso anterior. Resolvida as integrais e o sistema resultante, temos para ξ_0 e ξ_1:

$$e^{-\xi_0} = \frac{1}{\sqrt{2\pi\sigma^2}} \rightarrow \xi_0 = \frac{1}{2} ln(2\pi\sigma^2) \tag{3.114}$$

que é uma distribuição gaussiana ou normal, com média zero e variância σ^2.

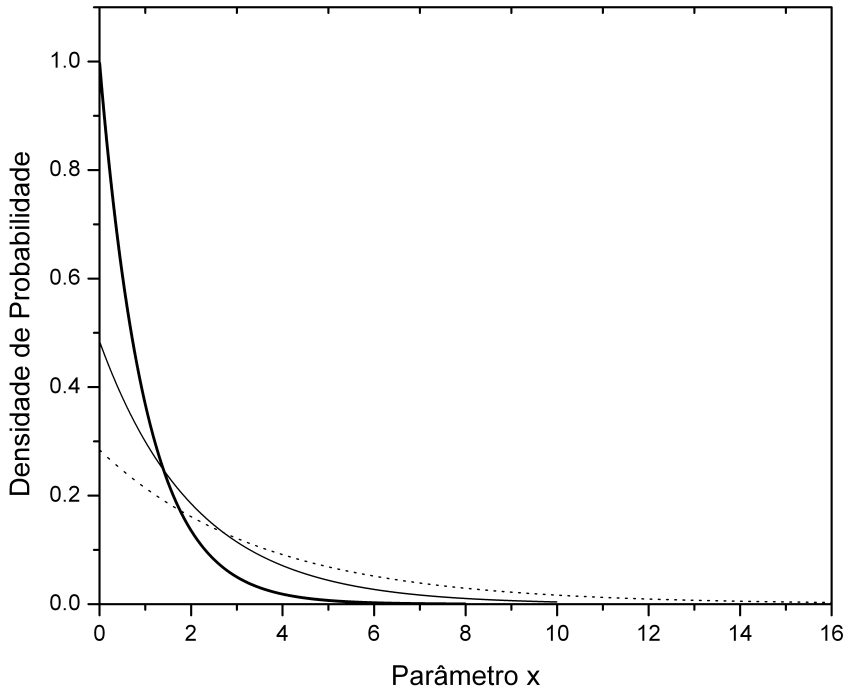

Figura 3.1: Gráficos da distribuição de máxima entropia obtida quando se apresenta como vínculo o valor esperado de x. O gráfico com a linha mais forte representa um vínculo de 1.0 para o valor da média, o tracejado 3.5 e a linha intermediária 2.0

Apresentamos na tabela 3.5 e gráfico (Fig. 3.2) os resultados do programa.

Tabela 3.3: Distribuição Normal, priori uniforme.

μ_1	μ_2	λ_0	λ_1	λ_2
0	0.01	-1.3836	0	50.0
0	0.1	-0.2324	0	5.0
0	0.5	0.5919	0	0.9506

Na tabela 3.4 mostramos o resultado de uma simulação na qual utilizamos uma distribuição a priori gaussiana de média zero e variância $\sigma^2 = 0.1$. Em seguida, encontramos uma distribuição

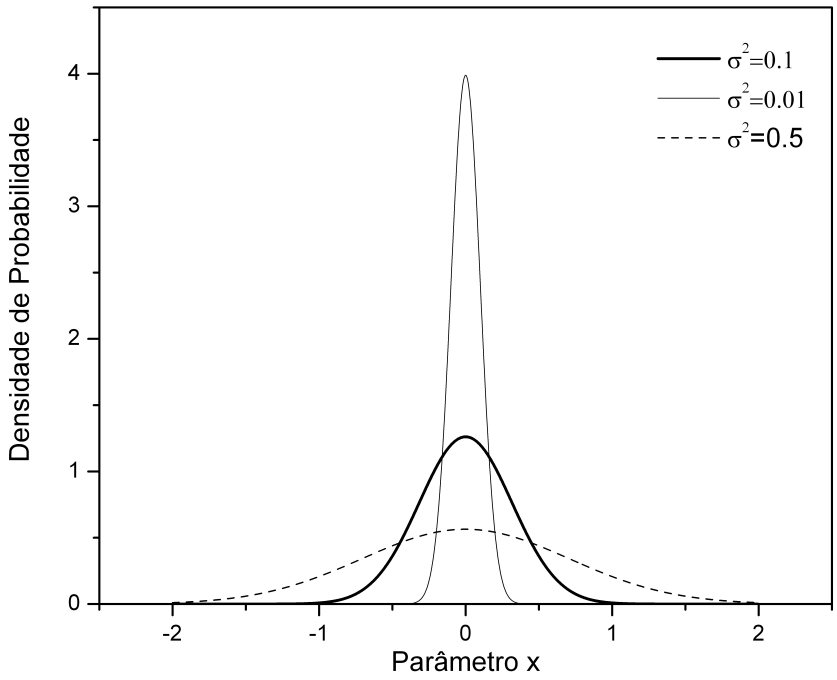

Figura 3.2: Distribuições normais geradas pelo método de máxima entropia quando se processa informações do tipo: valor esperado de x^2

posterior prescrevendo um $\mu_1 = E[x] = 1.0$. A Fig. 3.3 mostra a distribuição gerada. Note que obtemos uma distribuição do tipo gaussiana com uma média 1.0 e mesma variância da priori. Isto está de acordo com o princípio de máxima entropia no sentido de que atualizações somente são feitas quando há informação para tal.

Tabela 3.4: Distribuição Normal, com a distribuição a priori gaussiana de média zero e $\sigma^2 = 0.1$.

μ_1	λ_0	λ_1
1.0	48.1164	-90.0

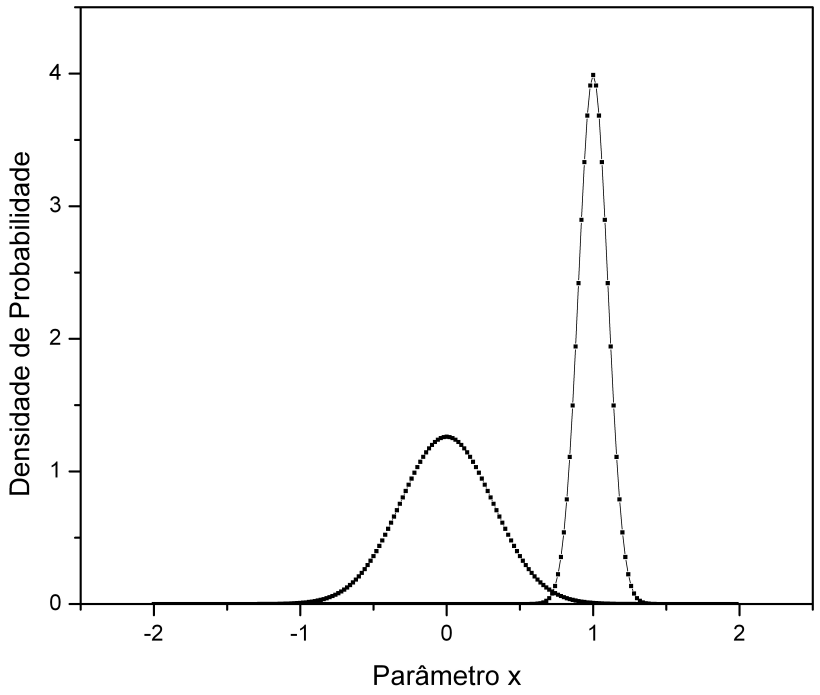

Figura 3.3: Distribuições a priori e posterior obtidas pelo método de máxima entropia; para gerar a distribuição posterior foi atualizado apenas o vínculo relacionda ao valor esperado de x, mantendo inalterado o valor da variância.

3.9.3 Distribuição Gama

A distribuição gama pode ser considerada uma distribuição de máxima entropia, quando os vínculos forem dados por (e $m(x) = 1$):

$$\int_0^\infty p(x)dx = 1 \tag{3.115}$$

$$\int_0^\infty xp(x)dx = \mu_1 \tag{3.116}$$

$$\int_0^\infty lnxp(x)dx = \mu_2 \tag{3.117}$$

onde os momentos considerados são $E(x)$ e $E(lnx)$.

Isso pode ser facilmente verificado pois a distribuição gama pode ser escrita como:

$$p(x) = exp[-\xi_0 - \xi_1 x - \xi_2 lnx] \tag{3.118}$$

Com $\xi_0 = -ln\frac{\lambda^k}{\Gamma(k)}$, $\xi_1 = \lambda$ e $\xi_2 = -(k-1)$. O problema proposto então é, dados μ_1 e μ_2 determinar ξ_0, ξ_1 e ξ_2. O caso da distribuição gama apresenta uma relação analítica entre λ e k e a média $m = E(x)$ e a variância $E(x - m^2) = \sigma^2$ dada por: $m = k/\lambda$ e $\sigma^2 = k/\lambda^2$. Assim, com os valores de λ e k podemos encontrar os valores de m e σ^2 e comparar com os valores analíticos.

Tabela 3.5: Distribuição Gama.

μ_1	μ_2	$k-1$	λ	ξ_0	ξ_1	ξ_2	m	σ^2
0.2	-2.0	0.4235	7.1175	-2.9146	7.1176	-0.4235	0.2000	0.0281
0.3	-1.5	0.8371	6.1221	-3.3888	6.1221	-0.8371	0.3000	0.0490
0.5	-1.0	0.7265	3.3844	-2.2006	3.3844	-0.7265	0.5101	0.1501

Pode-se verificar aqui, fazendo as comparações entre as médias e variâncias e os valores de k e λ, que o algoritmo produz bons resultados.

Diversas outras distribuições conhecidas podem ser obtidas pela aplicação do método, pela combinação de vínculos tais como: $E[x^\alpha]$, $E[ln(1+x)]$, $E[e^{-\alpha x}]$, dentre outras. Mostramos na tabela 3.6 algumas outras distribuições usuais, e quais vínculos a caracterizam.

Portanto, vimos como gerar algumas distribuições de probabilidade dados valores médios de funções. É importante salientar que a escolha dos limites em que a variável x pode assumir também é importante para a derivação da distribuição. Nos exemplos anteriores, os domínios da exponencial e da gama foram de $(0, \infty)$, enquanto para a distribuição normal foi $(-\infty, \infty)$.

3.10 Aplicação do Método de Máxima Entropia em Distribuições Simuladas

Na seção anterior, mostramos um método computacional para geração de distribuições de máxima entropia. Aplicamos esse algoritmo em alguns casos básicos, produzindo algumas dis-

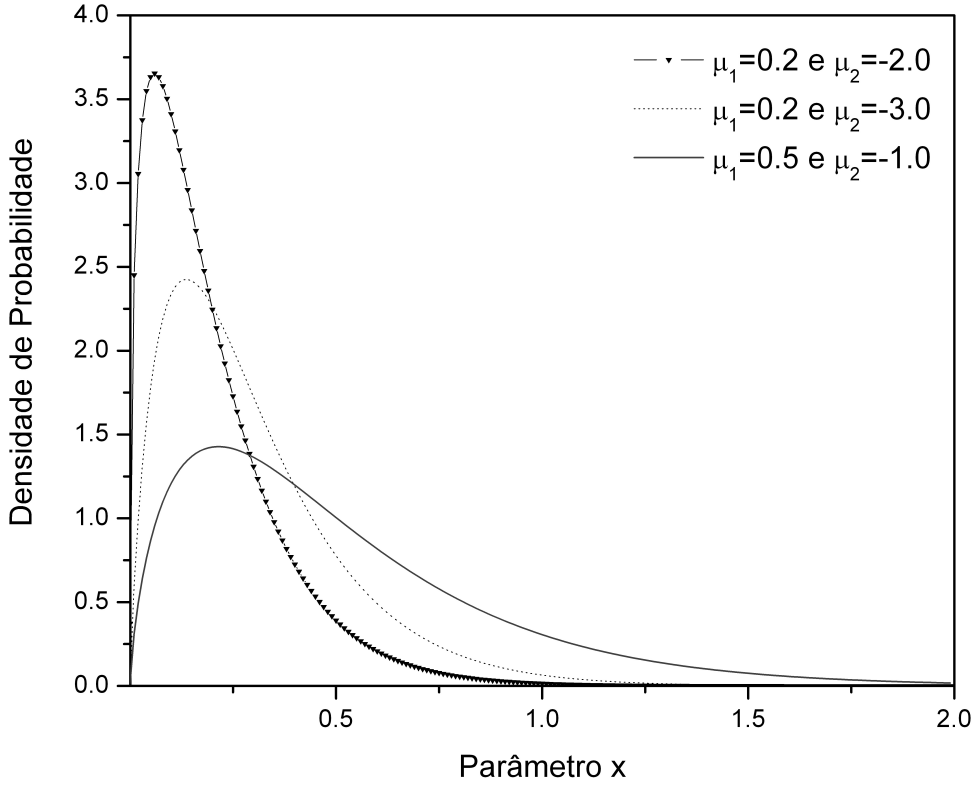

Figura 3.4: Distribuições gama gerados a partir dos vínculos: $E[x]$ e $E[lnx]$

tribuições canônicas, tais como *distribuição exponencial*, *distribuição normal* e *distribuição gama*; em todos os casos, os valores médios das funções de x utilizados foram arbitrários.

Aqui, vamos utilizar valores médios calculados a partir de distribuições simuladas, especificamente para o caso gama, ou seja, geramos distribuições gama com diferentes valores para seus parâmetros (k e λ) e, a partir disso, calculamos os valores de $E[x]$ e $E[ln(x)]$. Em seguida, usamos estes últimos valores como parâmetros de entrada do programa, gerando assim a distribuição de máxima entropia.

No gráfico (3.5) mostramos o resultado de nossos cálculos. Geramos quatro distribuições gama, respectivamente com os seguintes parâmetros: $k = 3.0$ e $\lambda = 5.0$, $k = 5.0$ e $\lambda = 10.0$, $k = 10.0$ e $\lambda = 1.0$, $k = 2.0$ e $\lambda = 2.0$, utilizando para efeito de comparação as distribuições calculadas analiticamente. Em todos estes casos, as distribuições simuladas foram geradas com 10000 pontos.

Tabela 3.6: Algumas distribuições geradas usando Máxima Entropia.

Nome	**Distribuição**	**Momentos**
Exponencial	$p(x) = \frac{1}{m}e^{-x/m}$, $x, m > 0$	$E[x] = m$
Normal	$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-x_2^2\sigma^2}$, $-\infty > x < \infty$, $\sigma > 0$	$E[x^2] = \sigma^2$
Gama	$p(x) = \frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x}$, $x, \lambda, k > 0$	$E[x]$,$E[lnx]$
Uniforme	$p(x) = \frac{1}{\beta-\alpha}$, $\alpha < x < \beta$	Normalização
Laplace	$p(x) = \frac{1}{2\sigma}e^{-\lvert x\rvert/\sigma}$, $-\infty < x, \infty$, $\sigma > 0$	$E[\lvert x\rvert]$
Cauchy	$p(x) = \frac{1}{\pi}\frac{1}{1+x^2}$,$-\infty < x < \infty$	$E[ln(1+x^2)]$
Logística	$p(x) = \frac{e^{-x}}{(1+e^{-x})^2}$	$E[x]$,$E[ln(1+e^{-x})]$

Como pode ser visto, em todas as curvas geradas o método recuperou a distribuição original quase perfeitamente para esse número de pontos simulados. Pode-se dizer que há uma relação unívoca entre os valores médios e a distribuição gerada, ou seja, um conjunto de vínculos especifica completamente uma determinada distribuição.

Em seguida, fixamos como parâmetros da distribuição o valor de $k = 5.0$ e $\lambda = 10.0$, variando o número de pontos gerados de 5 até os 10000 pontos já gerados. A partir desses pontos calculamos para cada curva os valores de $E[x]$ e $E[ln(x)]$, comparando com a curva analítica. As distribuições produzidas são mostradas nos gráficos da figura (3.6).

Para distribuições geradas com muitos pontos, as distribuições de máxima entropia estão em bom acordo com a analítica. Isso indica que podemos ter aqui um bom método para inferir valores de parâmetros quando há pouca informação disponível no problema. Note que, para 5 pontos gerados, apesar de a distribuição de máxima entropia não concordar com a analítica, ela já tem a forma de uma distribuição gama, o que é um bom sinal, visto o valor extremamente reduzido de pontos para análise.

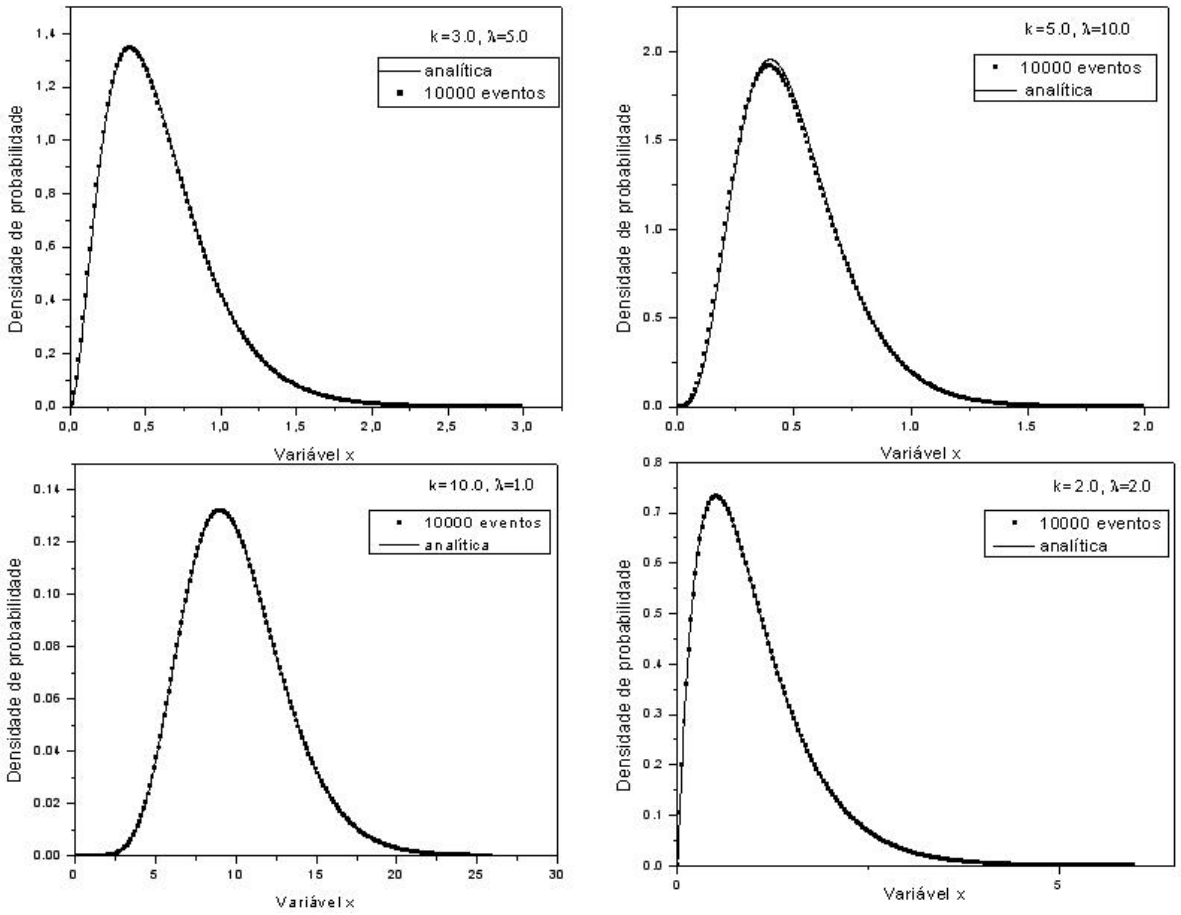

Figura 3.5: Comparação entre disribuições gama geradas a partir do método de Máxima Entropia e distribuições analiticas

3.11 Aplicação em Simulações de Raios Cósmicos

Agora, utilizamo-nos de uma simulação baseada no CORSIKA (*COsmic Ray SImulation for KAscade*), que é um programa baseado em cálculos de Monte Carlo para estudar a evolução de chuveiros de raios cósmicos pela atmosfera.

Quando atravessa a atmosfera, a partícula primária de alta energia vinda do espaço colide com moléculas de ar gerando outras partículas secundárias, em um processo em cadeia, até que, conforme a energia das partículas produzidas fica menor, a probabilidade de interação da partícula com o meio iguala-se com a probabilidade de absorção por átomos ionizados no ar, o que gera o fim do

chuveiro.

Na figura (3.7), mostramos o desenvolvimento longitudinal de 10 chuveiros de energia do primário de $10^5 GeV$ simulado pelo CORSIKA, usando o modelo hadrônico QGSjet e energias limiares de $3MeV$ para a componente eleltromagnética e $0,3GeV$ para a componente hadrônica. Esta curva representa o número de partículas em função da profundidade atmosférica.

Um modelo usado atualmente para ajustar os dados do desenvolvimento longitudinal das partículas foi feito originalmente por T. Gaisser e M. Hillas [27] e se baseia na seguinte função:

$$N(x) = N_{max}\left(\frac{x - x_0}{x_{max} - x_0}\right)^{\frac{x_{max}-x}{a+bx+cx^2}} \tag{3.119}$$

Ela fornece uma relação entre o número de partículas (N) e a profundidade atmosférica do chuveiro, em função de N_{max} que é o máximo número de partículas e x_{max} que é a profundidade em que temos o máximo número de partículas. Além desses, há outros 4 parâmetros ajustáveis livremente, o que torna o ajuste dessa curva bastante preciso.

Ao invés de ajustar a curva por Gaisser-Hillas, aplicamos o método de máxima entropia usando como informação disponível o valor médio de x, $[x] = 539,69$ e o valor médio de lnx, $[lnx] = 6,2192$. Pode-se verificar a partir da observação do gráfico que a curva dada pela distribuição de máxima entropia, que tem 2 parâmetros somente, se ajusta bem à curva simulada pelo CORSIKA.

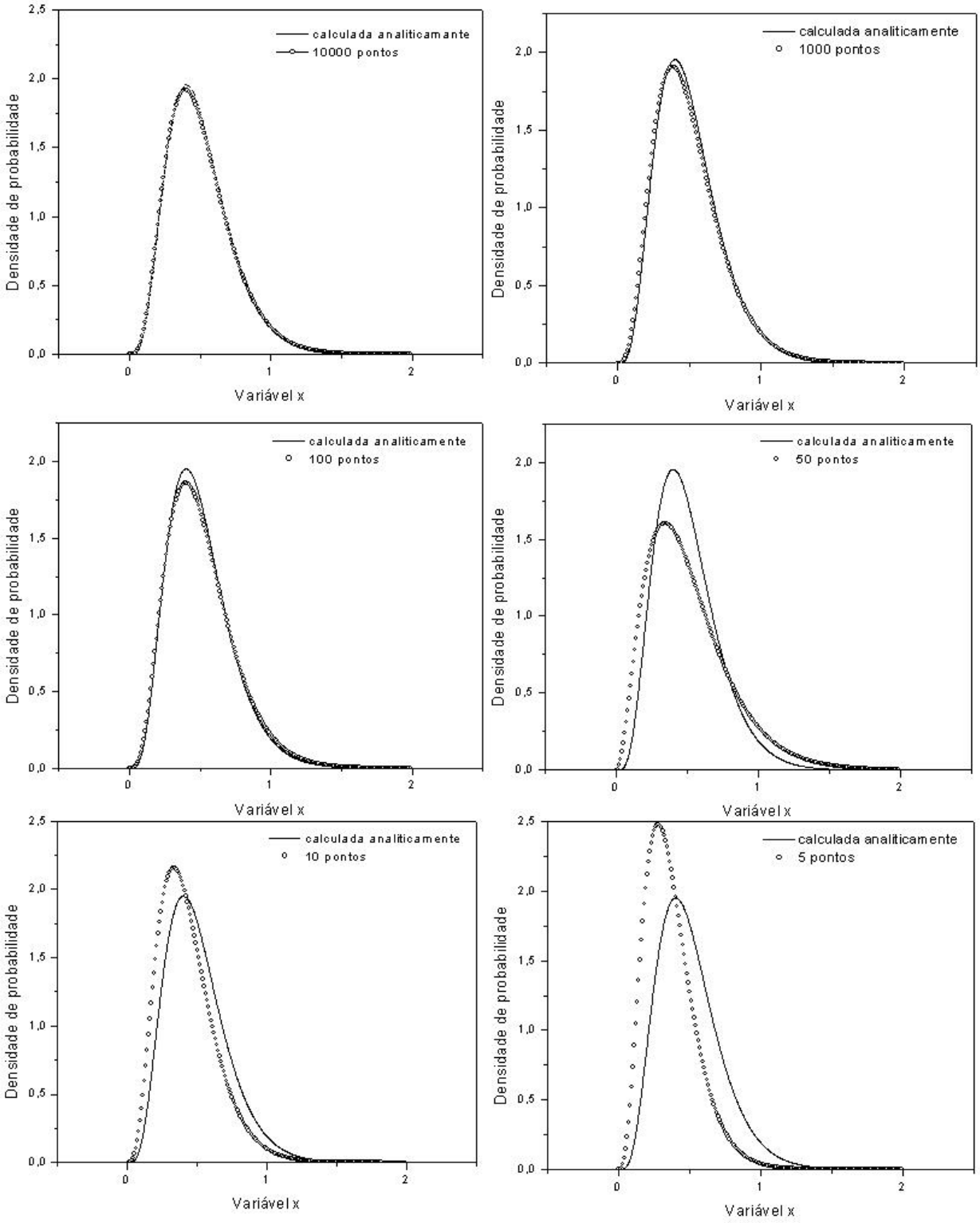

Figura 3.6: Na figura, mostramos a evolução das distribuições gama geradas por máxima entropia em função do número de pontos gerados; para efeito de comparação, em cada gráfico há a distribuição analítica correspondente

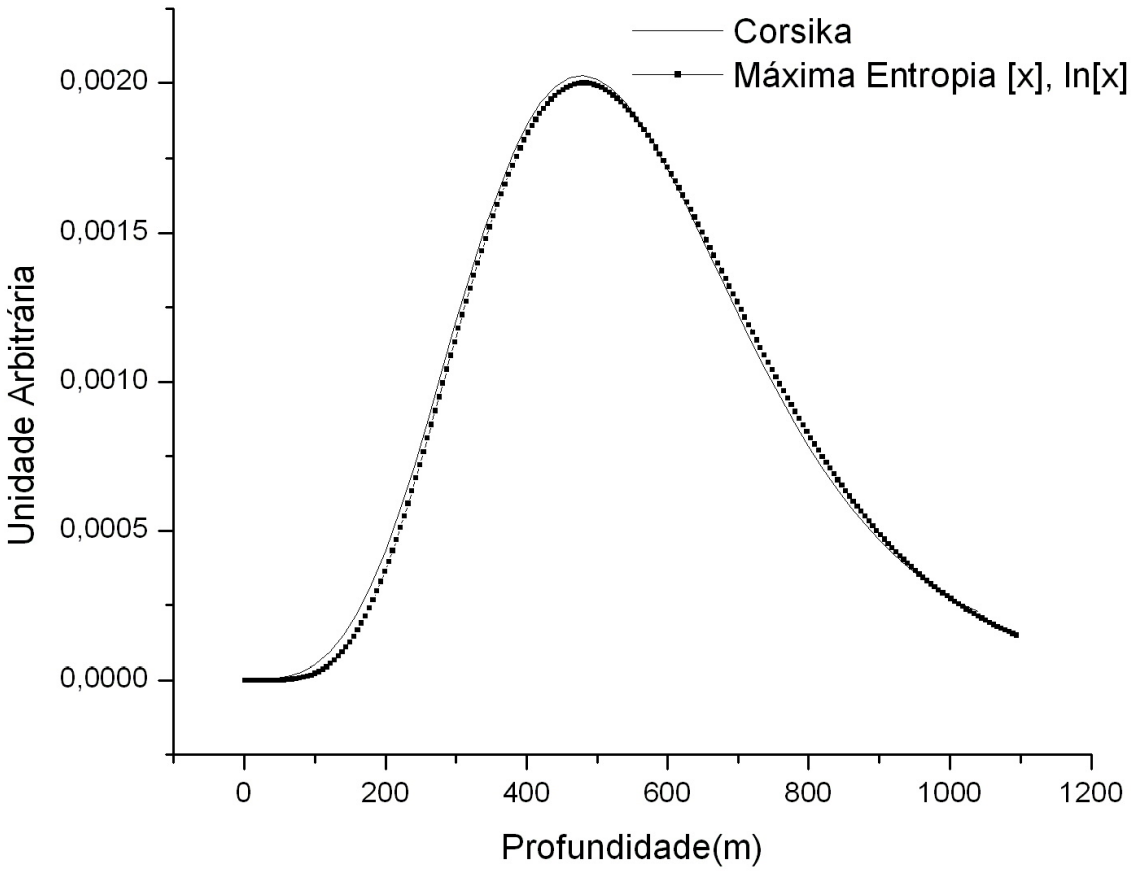

Figura 3.7: Gráfico da distribuição obtida por máxima entropia, partindo dos dados das simulações geradas pelo Corsika

4

Conclusões

Na presente dissertação, acreditamos ter apresentado de modo convincente dois métodos de inferência: a estatística bayesiana e o de máxima entropia, bem como indicado os procedimentos para o uso, em casos gerais. Como exposto anteriormente, o objetivo aqui não foi o de aplicar os métodos em muitos casos, mas sim ter mostrado alguns casos clássicos em que o procedimento dá muito bons resultados, o que acreditamos que tenha sido realizado.

Mostramos as condições em que a aplicação do método de Bayes torna-se interessante e útil, apresentando alguns exemplos ilustrativos . Demos nossa própria contribuição ao estudo do método, aplicando-o em um caso de interesse físico, e em específico, na física de altas energias: o modelo gama. Seguindo os procedimentos simples do método de Bayes inferimos os parâmetros de uma distribuição gama baseados em uma simulação.

Apresentamos o método de máxima entropia, expomos como ocorreu o a evolução do conceito de entropia, desde sua definição na termodinâmica clássica até as visões atuais, onde ela não está ligada a interpretações de calor, desordem ou incerteza mas é um instrumento de inferência.

Aplicando o método a casos clássicos, mostramos como distribuições de probabilidade com relevância em mecânica estatística podem ser derivadas, como é o caso das distribuições de Maxwell-Boltzmann, Bose-Einstein e Fermi-Dirac. A partir de um código fortran, apresentamos um procedimento que calcula numericamente distribuições de máxima entropia, e mostramos alguns resultados de algumas distribuições comuns: exponencial, normal e gama. Esses exemplos foram escolhidos porque neles podemos ver como as informações em forma de momentos determinam, dentro dos

domínios da variável aleatória e dos parâmetros, esses formatos de distribuição, sem qualquer outra hipótese matemática ou física.

É importante ressaltar aqui algumas questões do ponto de vista prático. O procedimento de aplicação da equação de Bayes

$$P(H_j|E_i, I) = \frac{P(E_i|H_j, I)P(H_j|I)}{P(E_i|I)}$$

requer a escolha de uma probabilidade a priori (Sec.2.6), bem como de uma função de verossimilhança. Outro ponto a se destacar é a regra de marginalização (sec 2.4), que permite isolarmos a probabilidade de ocorrência de um determinado parâmetro, enquanto os outros podem assumir qualquer valor - esse procedimento torna-se bastante útil para minimizar os efeitos de ruidos, num problema qualquer.

Quanto ao princípio de maximização da entropia, é importante salientar seu algoritmo básico de aplicação, que consiste em, a partir de vínculos dados por equações tais como

$$E[\phi_n(x)] = \int \phi_n(x)p(x)dx = \mu_n$$

atribuir uma distribuição de probabilidades do tipo

$$p(x) = m(x)exp\left[-\sum_{n=0}^{N}\xi_n\phi_n(x)\right]$$

em que o termo $m(x)$ contem informações a priori e pode ser igual a 1.

Deve-se notar, contudo, que os momentos $E[\phi_n(x)]$ não são escolhidos aleatoriamente, mas selecionados cuidadosamente a partir de experiências passadas que indiquem que a informação seja relevante.

A distribuição resultante é dada em termos de multiplicadores de Lagrange. A questão então se resume na determinação de tais multiplicadores

$$G_n(\vec{\xi}) = \int m(x)\phi_n(x)exp\left[-\sum_{n=0}^{N}\xi_n\phi_n(x)\right]dx = \mu_n$$

Temos então um conjunto de equações transcendentais que pode ser resolvido de diversas formas, sendo que o método de Newton é o canônico; este consiste em expandir as funções $G_n(\vec{\xi})$ em série de Taylor

$$G_n(\vec{\xi}) \cong G_n(\vec{\xi_0}) + (\vec{\xi} - \vec{\xi_0})^t [\mathbf{Grad} G_n(\vec{\xi})]_{(\vec{\xi}=\vec{\xi_0})} = \mu_n$$

Referências Bibliográficas

[1] E. D. Feigelson e G. J. Babu, Statistical Challenges in Modern Astronomy I-III, *Springer, New York, 1992, 1997, 2002*

[2] G. D'Agostini, Bayesian Reasoning in High Energy Physics. Principles and applications, *CERN Yellow Report* 99-03 (1999).

[3] T. Schwarz-Selinger, R. Preuss, V. Dose e W. von der Linden, Analysis of Multicomponent Mass Spectra Applying Bayesian Probability Theory, *J. Mass Spectrom*, **36**, 866 (2001).

[4] U. V. Toussaint, R. Fischer, K. Krieger e V. Dose, Depth Profile Determination with Confidence Intervals from Rutherford Backscattering Data, *New J. Phys.*, **1**, 11 (1999)

[5] G. L. Bretthorst, Bayesian Spectrum Analysis and Parameter Estimation, Notes in Physics, *Springer-Verlag, Berlin, 1988*, vol. 48; *J. Magn. Reson* **88**, 552 (1990)

[6] A. Caticha, Relative Entropy and Inference , in *Bayesian Inference and Maximum Entropy Methods in Sciences and Engineering*, **G.J. Erickson and Y. Zhai**, AIP Conf. Proc. **707** (2004)

[7] J. Bernoulli, *Ars Conjectandi.* Thurnisiorum, Basel (1713).

[8] T. Bayes, An essay towards solving a problem in the doctrine of chances, *Phil. Trans. Roy. Soc*, **53** (1763.)

[9] P. S. de Laplace, *Théorie Analytique des Probabilités* **Courcier Imprimeur**, Paris (1812).

[10] H. Jeffreys, *Theory of Probability* **Clarendon Press**, Oxford, (1939).

[11] R. T. Cox, *Am. J. Physics* **14** (1946).

[12] C. E. Shannon, A Mathematical Theory of Comunnication, *Bell Sys. Tech. J.* **27** (1948).

[13] E. T. Jaynes, Information Theory and Statistical Mechanics, *Phys. Rev.* **106** (1957)

[14] J. E. Shore e R. W. Johnson, Axiomatic Derivation of The priciple of Maximum Entropy and The Principle of Cross-Entropy, *IEEE Trans. In. Theory* **IT-26**, 26 (1981)

[15] G. D'Agostini, Bayesian Inference Processing Experimental Data Principles and Basic Applications, *Rep. Prog. Phys.* **66** 1383-1419 (2003).

[16] D. S. Sivia, *Data Analysis: A Bayesian Tutorial* **1ed. Oxford Science Publications**, Great Britain (1996).

[17] S.F. Gull, Bayesian inductive inference amd maximum entropy.,In *Maximum entropy and Bayesian methods in science and engineering*,**G.J. Erickson and C.R. Smith**, Kluwer, Dordrecht (1988).

[18] T. J. Loredo e D. Q. Lamb, Bayesian Analysis of Neutrino Observed from Supernova SN 1987A *Phys. Rev.* **D65** 063002, (2002).

[19] M. V. John e J. V. Narlikar, Comparison of Cosmological Models Using Bayesian Theory, *Phys. Rev.*, **D65**, 43506, (2002).

[20] P. Astone et al, Search fo Correlation between GRB's Detected by BeppoSAX and Gravitacional Wave Detectors EXPLORER and NAUTILUS, *Phys. Rev* **66**, 102002, (2002).

[21] A. I. Kinchin, Mathematical Foundations of Information Theory, *Dover Publ*, New York (1957)

[22] J. Skilling, The Axioms of Maximum Entropy, In *Maximum entropy and Bayesian Methods on Science and Engineering*,**G.J. Erickson and C.R. Smith**, Kluwer, Dordrecht (1988).

[23] J. N. Kapur, K. H. Kesavan, Entropy Optimization: Principles with Applications, *Academic*, Boston (1992)

[24] A. Caticha, Maximum Entropy and Bayesian Analysis: Entropic Prior Distributions, *Phys. Rev.* , **E70**, 046127 (2004)

[25] A. Caticha, A. Giffin, Updating Probabilities, *arXiv:physics/0608185 v1*, 26^{th} International Workshop on Bayesian Inference anda Maximum Entropy Methods (2006).

[26] A. D. Woodbury, A Fortran Program to Produce Minimum Relative Entropy Distributions, *Computers and Geosciences* **30**(2004).

[27] T. K. Gaisser and A. M. Hillas, Proc. 15th Int. Cosmic Ray Conf. (Plovdiv), 8 353, (1977).