Capítulo 1

Inferência Bayesiana, Simulação Estatística e suas Aplicações em Sistemas de Comunicações

1.1 Introdução

Frequentemente precisamos tomar decisões em situações envolvendo incertezas. Questões das mais prosaicas, como se devemos sair de casa com o guarda-chuva, até as mais complexas, como a profissão a ser seguida, nos compele a estimar a probabilidade de eventos futuros, avaliar riscos e fazer predições. Inferência estatística é a área da matemática aplicada que nos ajuda a realizar essas tarefas de forma mais rigorosa, complementando a intuição e o senso comum.

Duas vertentes da estatística, frequentista e bayesiana [1], disputam a primazia dos métodos de inferência e a análise de dados. A diferença fundamental entre as escolas reside na interpretação do conceito de probabilidade e por conseguinte no tratamento às quantias que se pretende inferir – os parâmetros, no jargão estatístico. Os frequentistas entendem a probabilidade como o limite da frequência relativa de ocorrência de um evento num universo de possibilidades, e portanto restringem a análise probabilística a eventos reprodutíveis. Em oposição, bayesianos se alicerçam na visão da probabilidade como uma medida para o grau de incerteza subjetiva sobre possíveis eventos.

Como consequência, os frequentistas sustentam que os parâmetros sejam tratados como constantes desconhecidas, enquanto os bayesianos os interpretam como variáveis aleatórias, ainda que sejam pré-definidos. Como esses últimos interpretam a probabilidade como medida do grau de incerteza, o desconhecimento sobre os parâmetros é motivo legítimo para tratá-los como variáveis aleatórias com suas distribuições de probabilidade associadas.

Ilustremos essas diferenças com o familiar experimento de lançamento de um dado. Deseja-se calcular, por exemplo, a probabilidade de que seja primo o número na face voltada para cima. Nos primórdios da teoria da probabilidade, explorava-se a simetria do experimento para obter as probabilidades elementares de cada possível resultado em um lançamento, e usavam-se certas regras para se calcular a probabilidade

de um evento composto. Por simetria, os seis resultados devem ser igualmente prováveis, e como é forçoso que a soma das probabilidades seja unitária, segue que a probabilidade de cada face é 1/6. A probabilidade de o resultado ser um número primo seria então igual à razão entre a quantidade de números primos entre 1 e 6 (2,3 e 5) e a quantidade total resultados possíveis, ou seja 1/2.

Não obstante seu apelo intuitivo, esse argumento seria rejeitado por um frequentista. É impossível determinar previamente as probabilidades envolvidas, pois essas são o resultado da frequência relativa de cada evento após muitos experimentos. A probabilidade de sair um número primo é obtida contando-se as ocorrências em que os números 2,3 ou 5 foram efetivamente obtidos numa longa sequência de lançamentos, e dividindo-se o resultado pela quantidade total de lançamentos. Espera-se que essa razão coincida com a obtida pela solução anterior se o dado não for viciado. No entanto, um frequentista só poderia julgar um dado justo (ou não) após a realização dos experimentos.

Um bayesiano posiciona-se entre esses extremos: considerações de simetria importam, porém variações são permitidas à medida que são realizados novos experimentos. Antes do primeiro lançamento, as seis faces seriam consideradas igualmente prováveis; a cada novo experimento, as probabilidades poderiam variar num sentido sentido ou outro, dependendo da sequência de resultados obtidos até o momento. A probabilidade é portanto, uma quantidade dinâmica, que parte de um valor plausível e torna-se progressivamente mais precisa com a aquisição de novas evidências. No limite, após infinitas tentativas, o bayesiano e o frequentista tendem a remover suas divergências.

A escola bayesiana tem uma longa e fascinante história [2]. Foi dada à luz em 1740 pelo então desconhecido ministro presbiteriano britânico chamado Thomas Bayes, descobridor da regra de atualização de probabilidade que leva o seu nome. Essa regra é uma enganosamente simples equação que relaciona a probabilidade de uma hipótese (ou evento) após experimentos serem realizados com a probabilidade prévia – antes dos dados serem observados – e uma função que mede o quão surpreendente são as observações para o caso de ser verdadeira cada uma das possíveis hipóteses. Negligenciado durante a vida de Bayes, o teorema foi redescoberto independentemente três décadas mais tarde por Pierre Laplace, que lhe deu uma nova formulação e o aplicou em inúmeros problemas práticos da época.

Condenada por estatísticos do calibre de Fisher e Pearson como anti-científica devido a sua natureza assumidamente subjetiva, a visão bayesiana continuou sendo adotada, ainda que de forma marginal, durante a maior parte do século XX, pois fornecia soluções de problemas para os quais a visão frequentista falhava. Após décadas de ostracismo, avanços teóricos feitos por matemáticos como Jeffrey, De Finet e Savage tornaram o bayesianismo mais respeitável. Todavia, sua aceitação definitiva e popularização só vieram no fim da década de 80, quando o surgimento de métodos numéricos baseados em MCMC (Markov Chain Monte Carlo) permitiram a aplicação da ferramenta em inúmeros problemas até então intratáveis por requerer intenso poder computacional. Atualmente, métodos bayesianos estão no cerne de uma miríade de aplicações como aprendizado por máquina, genética, filtragem anti-*spam*, investigação

forense e predição de resultados de eleições.

Em problemas de engenharia, incluindo telecomunicações, soluções mais efetivas são obtidas explorando a maior flexibilidade do paradigma bayesiano frente à mais tradicional visão frequentista. A incorporação de qualquer conhecimento prévio do engenheiro – muitas vezes subjetivo e difícil de quantificar – sobre as possíveis soluções para o problema em questão, é de grande valia em problemas reais em que dados objetivos são escassos. Em um sistema de recepção em comunicações sem-fio, o conhecimento sobre o comportamento típico do canal de comunicação para o ambiente em questão, além das propriedades estatísticas dos sinais transmitidos, podem ser incorporados na solução e permitir deteção mais acurada de sinais. O tratamento de parâmetros como variáveis aleatórias, apanágio da escola bayesiana, permite a marginalização das variáveis que não são de interesse imediato, levando a soluções ótimas para os parâmetros que realmente deseja-se estimar – por exemplo, o sinal que foi enviado.

Este capítulo se destina à apresentação dos rudimentos de inferência bayesiana e das técnicas de simulação estatística que permitem sua aplicação a problemas complexos. Como fio condutor da exposição, consideraremos o problema de recepção de sinais em comunicações sem-fio, em que o canal variante no tempo e sujeito a ruído tornam-o particularmente adequado a ferramentas bayesianas. As primeiras seções dirão respeito aos fundamentos da análise de dados Bayesianas, e incluirão uma exposição sobre o teorema de Bayes, a formalização do conceito de distribuição *a priori* – em particular, distribuição *a priori* conjugada e priori não-informativa –, modelo Bayesiano hierárquico e eliminação de parâmetros auxiliares. Nas últimas seções, discorreremos sobre técnicas numéricas de exploração de distribuições baseadas em Cadeias de Markov e Monte Carlo (MCMC), que permitem resolver problemas de inferência que seriam de outro modo intratáveis. O foco será nos populares e poderosos métodos da classe MCMC, em especial o algoritmo de Metropolis-Hastings e o amostrador de Gibbs. Terminamos o capítulo aplicando as técnicas apresentadas ao problema de deteção de sinais em comunicações sem-fio para um canal seletivo em frequência e sujeito a ruído aditivo gaussiano.

1.2 O Teorema de Bayes

Usamos inferência estatística quando desejamos obter informações sobre grandezas ocultas a partir de dados observáveis que podem ser descritos por modelos probabilísticos. Num sistema CDMA (Code Division Multiplex Access), por exemplo, diversos usuários compartilham o mesmo meio, transmitindo na mesma faixa de frequência, e sujeitos às distorções e ruído do canal. É tarefa do receptor analisar essa complicada combinação de sinais – que dependem apenas indiretamente das mensagens transmitidas – e inferir as informações que mais provavelmente foram enviadas por cada usuário.

Vamos formalizar o conceito de inferência especificando uma variável $\mathbf{x}$ possivelmente multidimensional cuja distribuição é descrita por $p(\mathbf{x};\boldsymbol{\theta})$. Aqui, $\boldsymbol{\theta}$ reúne as quantias desconhecidas que se deseja estimar

e será referenciada como o conjunto de parâmetros – embora por vezes, como no exemplo do canal de comunicação, esses parâmetros sejam também dados.

Como vimos, a escola frequentista entende os parâmetros $\boldsymbol{\theta}$ como valores fixos e desconhecidos. Consequentemente, a inferência é feita com base na distribuição $p(\mathbf{x}|\boldsymbol{\theta})$, conhecida como verossimilhança, e entendida como a distribuição de probabilidade dos dados observados para um certo conjunto de parâmetros $\boldsymbol{\theta}$. Em contraste, a escola bayesiana enxerga os parâmetros como variáveis aleatórias com certas distribuições associadas. Antes de os dados serem observados, a distribuição associada aos parâmetros é chamada de distribuição *a priori* (ou simplesmente priori), $p(\boldsymbol{\theta})$, e quantifica o grau de confiança do projetista sobre os possíveis valores dos parâmetros. Após os dados serem levados em consideração, a nova distribuição dos parâmetros é conhecida como distribuição *a posteriori* (ou simplesmente posteriori), e condensa a combinação do conhecimento prévio com as informações recém adquiridas.

A relação entre essas grandezas é definida pelo teorema (ou regra) de Bayes. Através do teorema, podemos obter a distribuição *a posteriori* a partir da distribuição *a priori*, levando em conta os dados observados. Seja $p(\boldsymbol{\theta}|H)$ a distribuição *a priori*, em que H (de história) representa todas as informações à disposição do usuário sobre os parâmetros de interesse antes de os dados estarem disponíveis. O teorema estabelece que a distribuição *a posteriori* é obtida essencialmente pelo produto da distribuição *a priori* com a verossimilhança:

$$p(\boldsymbol{\theta}|\mathbf{x}, H) = \frac{p(\mathbf{x}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H)}{p(\mathbf{x}|H)}. \tag{1.1}$$

Nessa expressão $p(\mathbf{x})$ é a densidade de probabilidade associada ao vetor de dados $\mathbf{x}$ que leva em conta todas as possibilidades de $\boldsymbol{\theta}$. Como $\mathbf{x}$ é conhecido, $p(\mathbf{x})$ é apenas uma constante de proporcionalidade que garante que seja unitária a integral do quociente na região definida pelo espaço amostral dos parâmetros. Sendo assim, $p(\mathbf{x})$ é dado por:

$$p(\mathbf{x}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{1.2}$$

Sendo constante, $p(\mathbf{x}|H)$ não afeta forma da função $p(\boldsymbol{\theta}|\mathbf{x}, H)$. Por essa razão é usual remover $p(\mathbf{x})$ do teorema de Bayes e reescrevê-lo como uma relação de proporcionalidade:

$$p(\boldsymbol{\theta}|\mathbf{x}, H) \propto p(\mathbf{x}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H). \tag{1.3}$$

Podemos resumir o teorema informalmente como: a probabilidade posterior de um evento é o produto da probabilidade prévia com a probabilidade associada aos dados observados (evidência). Ou ainda:

$$\text{Probabilidade posterior} = \text{Verossimilhança} \times \text{Probabilidade prévia} \tag{1.4}$$

Sendo assim, o teorema permite ponderar o conhecimento prévio com as novas informações após a realização da experiência, gerando conhecimento novo e melhorado.

Exemplo 1: exame médico

Vamos ilustrar o teorema numa situação de importância prática que infelizmente confunde tanto leigos quanto especialistas. Apresentaremos aqui uma versão resumida do problema exposto em [2]. Uma mulher recebeu o resultado de sua mamografia e avaliação foi anormal. Qual a probabilidade de que ela tenha câncer de mama? Para responder à pergunta precisamos de algumas estatísticas básicas, a saber:

1. Probabilidade de mamografia anormal na ausência da doença: 10 %.
2. Probabilidade de mamografia anormal na presença da doença: 80 %.
3. Probabilidade de uma mulher qualquer sofrer de câncer de mama: 0.4 %.

Muitos estariam inclinados a afirmar que a probabilidade de a mulher ter a doença é de 80 % – o percentual de vezes em que o teste resulta anormal quando a doença está presente. Um dado de importância fundamental costuma ser esquecido: a probabilidade prévia, isto é, *antes* de o exame ser realizado, de que uma mulher qualquer sofra de câncer de mama. Dada a raridade da doença (0.4 %), a evidência teria que ser muito forte para elevar significativamente a probabilidade em relação ao seu patamar inicial. Decerto esperamos que o resultado anormal aumente essa probabilidade inicial, mas devido à imperfeição do exame, podemos antecipar que a probabilidade posterior esteja num valor intermediário entre 0.4 e 80 %.

O teorema de Bayes permite calculá-la exatamente. Chamemos de A o evento "mulher tem câncer" e B o evento "mamografia anormal". Queremos calcular probabilidade *a posteriori* de a mulher estar doente sabendo do resultado anormal do exame, ou seja, $p(B|A)$, que Bayes nos fornece:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}. \tag{1.5}$$

Pelos dados do problema, temos:

1. $p(A|B) = 0.8$
2. $p(B) = 0.004$
3. $p(A) = p(A|B)p(B) + p(A|\overline{B})p(\overline{B}) = 0.8 \times 0.004 + 0.1 \times 0.996 = 0.1028.$

Nessa última expressão $p(A)$ é a probabilidade de o exame resultar anormal, obtido somando-se as probabilidades de resultado anormal quando a mulher tem a doença e quando ela está saudável. A barra acima do evento indica o seu complementar, no caso a sua negação.

A resposta final é obtida substituindo esses valores em 1.5

$$p(B|A) = \frac{0.8 \times 0.004}{0.1028} = 0.311 = 3.11\% \tag{1.6}$$

Esse resultado contra-intuitivo mostra a importância de se considerar as probabilidades prévias e recomenda que resultados de exames sejam interpretados cautelosamente. Muitas pessoas tem suas vidas devastadas ao receber resultado positivo para uma doença séria, como AIDS, apesar de que a raridade da doença e a imperfeição dos exames implicarem uma probabilidade não desprezível de que o vírus não esteja presente.

Exemplo 2: canal binário simétrico

A Fig. 1.1 diz respeito a um canal de comunicações em que transmitem-se símbolos binários (0 ou 1) e, devido a imperfeições do canal, há uma probabilidade p usualmente menor que 1 de que um símbolo seja corretamente detectado. Vamos supor que a probabilidade de o dígito 0 ser enviado seja p_0. Claramente $p_1 = 1 - p_0$.

Chamemos de X o símbolo enviado e de Y o recebido. É útil calcular a probabilidade inversa de que o símbolo X tenha sido enviado sabendo que Y foi recebido. Invoquemos Bayes novamente para nos ajudar.

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \tag{1.7}$$

em que $p(Y = 1) = p_1 p + p_0(1 - p)$ e $p(Y = 0) = p_0 p + p_1(1 - p)$

Assim,

$$p(X = 1|Y = 1) = \frac{pp_1}{p_1 p + p_0(1 - p)} \tag{1.8}$$

$$p(X = 0|Y = 0) = \frac{pp_0}{p_0 p + p_1(1 - p)} \tag{1.9}$$

Naturalmente a probabilidade de que haja erro de transmissão em cada caso é obtida calculando-se o complementar de cada uma das expressões acima.

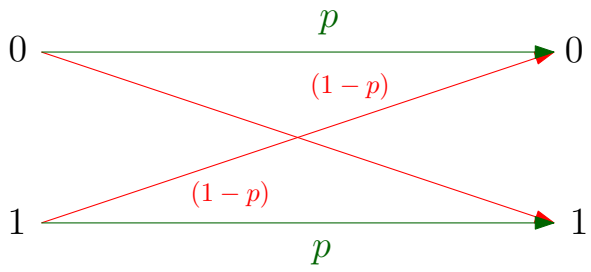

Figura 1.1: Canal binário simétrico. Símbolos binários são transmitidos com probabilidade de erro $(1-p)$

1.3 Bayesianos vs Frequentistas

A exposição detalhada das diferenças e similaridades entre as escolas bayesiana e frequentista está claramente além do escopo do capítulo. Aqui nos limitaremos a apresentar o que julgamos essencial sobre o assunto e exporemos um exemplo no qual as principais diferenças entre as escolas ficarão patentes.

No âmago do processo de inferência adotado por ambos frequentistas e Bayesianos está a *função de verossimilhança*, aqui denotada por $l(\mathbf{x};\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$. Essa função é entendida como a probabilidade de os dados $\mathbf{x}$ serem observados quando se assume que os parâmetros corretos são $\boldsymbol{\theta}$. Na escola frequentista a inferência é usualmente feita com base apenas na verossimilhança, e $\boldsymbol{\theta}$ é entendido como um parâmetro que aparece na distribuição de $p(\mathbf{x})$ e não como uma variável condicionante. Por sua vez, a escola Bayesiana trata a verossimilhança como o elemento de ligação entre a priori e a posteriori – através da regra de Bayes – e $\boldsymbol{\theta}$ é entendida como variável que condiciona a distribuição dos dados observados $\mathbf{x}$.

O critério de inferência de Máxima Verossimilhança (ML, do inglês *Maximum Likelihood* [3]) é uma das mais populares ferramentas de inferência do repertório frequentista. A ideia por trás do ML é bem intuitiva: o valor dos parâmetros estimado é aquele que torna os dados observáveis os mais prováveis. Ou, como o nome sugere, o valor de $\boldsymbol{\theta}$ estimado é aquele que maximiza a função de verossimilhança.

A escola bayesiana também permite obter estimativas pontuais dos parâmetros, mas os diversos possíveis critérios baseiam-se na distribuição *a posteriori* e não na verossimilhança. O Mínimo Erro Quadrático Bayesiano (BMSE, *Bayeian Minimum Square Error*) e o Máximo *A Posteriori* (MAP, *Maximum A Posteriori*) são duas das escolhas mais usuais. Pelo critério MAP o valor estimado para os parâmetros é aquele que maximiza a distribuição *a posteriori*, enquanto o BMSE resulta na média dessa distribuição.

Examplo 3: Lançamento de moeda

Vamos ilustrar essas diferenças com um problema corriqueiro e outro mais relevante em engenharia. Considere o familiar problema de se estimar a probabilidade θ de sair 'cara' no lançamento de uma moeda. Já vimos no problema do lançamento de um dado que considerações de simetria para determinação das probabilidades são controversas, não obstante seu apelo intuitivo. Confrontado por essa questão, um frequentista diria que experimentos precisam ser feitos para que uma resposta significativa seja fornecida. Suponha que ele lance a moeda três vezes e o resultado tenha sido 'coroa' nos três casos. Ele então calcula a verossimilhança $p(\mathbf{x}|\theta)$, isto é, a probabilidade de o resultado $\mathbf{x} = [1\ 1\ 1]$ – representando três coroas em três lançamentos – seja obtido considerando cada possível valor de $\theta \in [0\ 1]$. Supondo independência entre os lançamentos, a probabilidade de três coroas é o produto das probabilidades individuais, isto é, $p(\mathbf{x}|\theta) = (1-\theta)^3$. Claramente a verossimilhança é maximizada quando $\theta = 0$. O frequentista não teria muita confiança nessa estimativa de máxima verossimilhança porque poucos lançamentos foram feitos; mas se forçado a dar uma resposta ao problema ele diria ser nula a probabilidade de sair 'cara'.

O Bayesiano analisando o mesmo probblema também não daria uma resposta imediata. Ele começaria

definindo uma probabilidade *a priori* $p(\theta)$ refletindo sua experiência prévia sobre o lançamento de moedas. Embora não estritamente necessário, é razoável assumir que θ tem maior chance de estar em torno de 1/2 do que nos limites (0 ou 1). Uma possibilidade seria a curva em azul da Fig. 1.2 (a). Combinando essa priori com a verossimilhança ele obteria a posteriori (em verde na figura), cuja maximização forneceria para θ um valor à esquerda de 1/2. À medida que novos experimentos são feitos, posterioris cada vez mais informativas são geradas (Fig 1.2 (b)) e estimativas mais precisas são obtidas para θ.

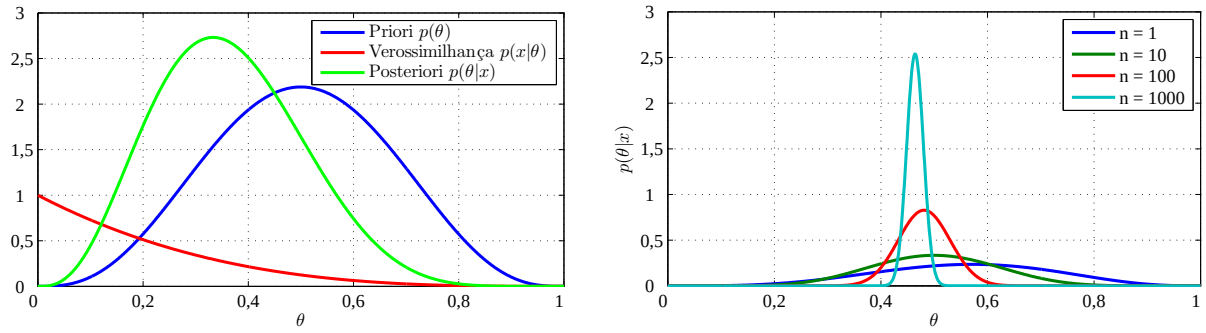

Figura 1.2: Determinando a probabilidade de sair 'cara' no lançamento de uma moeda. (a) Exemplo da priori, verossimilhança e posteriori após 3 lançamentos com resultado 'coroa'; o máximo da posteriori está deslocado para a esquerda de 1/2. (b) Posterioris ficam cada vez mais concentradas quando o número de lançamentos n aumenta.

Exemplo 4: Modelo linear generalizado

Consideremos agora um modelo de grande utilidade que aparece frequentemente em diversos problemas de engenharia elétrica, inclusive telecomunicações. O chamado modelo linear generalizado descreve uma situação em que os dados observados dependem linearmente de um conjunto de valores desconhecidos a serem estimados, e ruído aditivo está presente na medida. Bastante geral, o modelo pode descrever, por exemplo, um sinal formado por uma mistura de senóides com amplitudes desconhecidas imersas em ruído, ou o sinal recebido em comunicações sem fio com múltiplos percursos e ruído aditivo. Matematicamente o vetor de N elementos contendo os dados observados, $\mathbf{x}$, é a soma de um produto matricial e um ruído $\mathbf{v}$:

$$\mathbf{x} = \mathbf{G}\boldsymbol{\theta} + \mathbf{v} \tag{1.10}$$

onde o produto $\mathbf{G}\boldsymbol{\theta}$ determina a relação linear entre os dados $\mathbf{x}$ e os parâmetros desconhecidos $\boldsymbol{\theta}$.

Em muitas aplicações, $\mathbf{v}$ pode ser satisfatoriamente aproximado como ruído branco gaussiano de média zero e variância σ_v^2 (muitas vezes desconhecida). Se a matriz $\mathbf{G}$ for conhecida, a distribuição de $\mathbf{x}$ dado $\boldsymbol{\theta}$ é igual à distribuição de $\mathbf{v}$ com média deslocada de $\mathbf{G}\boldsymbol{\theta}$. Logo a verossimilhança é dada por:

$$p(\mathbf{x}|\boldsymbol{\theta}) = p_v(\mathbf{x} - \mathbf{G}\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_v^2}(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})\right\}. \tag{1.11}$$

A maximização dessa função em relação a $\boldsymbol{\theta}$ fornece o mesmo resultado que a minimização do argumento da exponencial. Igualando a zero o gradiente desse argumento, obtemos o estimador ML:

$$\boldsymbol{\theta}^{\mathrm{ML}} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{x}. \tag{1.12}$$

A escola bayesiana resolveria esse mesmo problema de forma diferente. Os parâmetros $\boldsymbol{\theta}$ seriam tratados como uma variáveis aleatórias e ganhariam uma distribuição *a priori*, refletindo o conhecimento prévio sobre os parâmetros. Considerando, para fins de ilustração, uma priori gaussiana com média $\mathbf{m}_{\boldsymbol{\theta}}$ e matriz de covariância $\mathbf{C}_{\boldsymbol{\theta}}$, teríamos uma posteriori cuja maximização geraria o seguinte estimador para $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{\mathrm{MAP}} = (\mathbf{G}^T\mathbf{G} + \sigma_v^2\mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1}(\mathbf{G}^T\mathbf{x} + \sigma_v^2\mathbf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{m}_{\boldsymbol{\theta}}). \tag{1.13}$$

Se os elementos de $\mathbf{C}_{\boldsymbol{\theta}}$ forem altos na equação acima, as parcelas introduzidas pela priori se tornam irrelevantes, e então a solução MAP se aproxima da ML. Isso é plausível, uma vez que valores altos para $\mathbf{C}_{\boldsymbol{\theta}}$ indicam uma priori bastante ampla, refletindo um conhecimento prévio vago sobre os parâmetros, o que implica maior importância dos dados observados no resultado final. Por outro lado, se muitos dados são observados, e consequentemente $\mathbf{x}$ é um vetor de dimensão alta, os termos dependendo de $\mathbf{G}$ e $\mathbf{x}$ dominariam os termos dependentes dos parâmetros da priori, que são fixos. Nesse caso, com muitos dados, o MAP também se aproximaria do ML, independentemente da priori.

Esses dois exemplos nos permitem concluir que as escolas bayesiana e frequentista tendem a concordar quando muitos dados são disponíveis, mas divergem em muitas situações práticas em que dados são escassos e o conhecimento prévio é relevante.

1.4 Elementos de Inferência Bayesiana

Vimos até aqui conceitos que são comuns às escolas Bayesianas e frequentistas, embora cada escola os usem de forma diferente. As seções seguintes dizem respeito a elementos que são exclusividade da escola bayesiana.

1.4.1 Modelo Bayesiano Hierárquico

Em muitos problemas práticos de engenharia, os dados observados dependem de diversos parâmetros que podem apresentar uma relação hierárquica entre si. Em comunicações sem fio, por exemplo, o sinal recebido depende do canal que pode ser caracterizado por um filtro linear com coeficientes desconhecidos. Nesse contexto, um bayesiano atribuiria um modelo probabilístico para os coeficientes do canal, e como os parâmetros desse modelo seriam também desconhecidos, o estatístico os descreveria probabilisticamente em termos de outros parâmetros – que passam a ser chamados de hiperparâmetros.

A Figura 1.3 ilustra esse conceito de forma mais geral. Possíveis conjuntos de dados observados, x_1, x_2, ..., x_n, dependem de parâmetros θ_1, θ_2, ..., θ_n, respectivamente, que por sua vez são instâncias de uma variável aleatória descrita pelo hiperparâmetro ϕ. Esse modelo poderia representar, por exemplo, as notas de alunos de n diferentes escolas, cada uma com uma determinada média θ_i $i \in \{1, \ldots, n\}$, que por sua vez depende de uma distribuição mais básica escrita em termos de ϕ (possivelmente a média geral dos alunos daquela região).

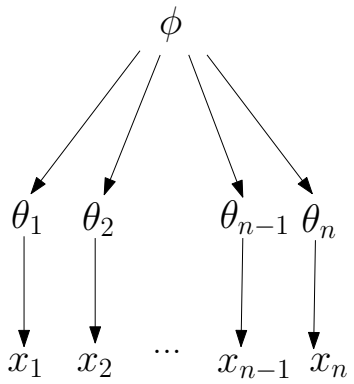

Figura 1.3: Exemplo de modelo bayesiano hierárquico. Os conjuntos de dados x_1, x_2, ..., x_n, dependem de parâmetros θ_1, θ_2, ..., θ_n, respectivamente, que por sua vez são instâncias de variáveis aleatórias descritas pelo parâmetro ϕ.

O teorema de Bayes fornece a distribuição *a posteriori* dos parâmetros desconhecidos em função dos dados observados. No exemplo em questão, a distribuição *a posteriori* é denotada por $p(\theta|x_1, \ldots, x_n)$, e a distribuição *a priori* de θ depende de ϕ, que por sua vez é caracterizado pela priori $p(\phi)$. Logo,

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})}{p(x_1, \ldots, x_n)}. \tag{1.14}$$

em que usamos $p(\theta, \phi) = p(\theta|\phi)p(\phi)$.

1.4.2 Eliminação de parâmetros

É comum o vetor de parâmetros conter elementos que não são de interesse para estimação. No exemplo do canal de comunicação, os coeficientes do canal são necessários apenas para termos um modelo completo do fenômeno em estudo; mas o foco do problema está na estimação dos dados enviados.

Por tratar parâmetros como variáveis aleatórias, a escola Bayesiana nos autoriza a trabalhar apenas com as variáveis de interesse, integrando os demais, os ditos parâmetros *nuisance*. Particionado $\boldsymbol{\theta}$ entre os parâmetros *nuisance* e os de interesse, de tal forma que $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\psi})$, a posteriori para os parâmetros de interesse calcula-se como:

$$p(\boldsymbol{\phi}) = \int_{\psi} p(\boldsymbol{\phi}, \boldsymbol{\psi}|\mathbf{x}) d\boldsymbol{\psi}. \tag{1.15}$$

Essa integral nem sempre pode ser feita analiticamente ou com técnicas clássicas de integração numé-

rica. A dificuldade é especialmente severa quando a distribuição é multivariável e não possui uma forma conhecida. Nesses casos, os algoritmos de MCMC discutidos mais adiante são especialmente recomendados.

1.4.3 Distribuição *a priori*

Como vimos, a distribuição *a priori* é uma função que reflete o conhecimento do usuário sobre os possíveis valores dos parâmetros antes de levar os dados em conta. Embora critérios frequentistas possam guiar a escolha de $p(\boldsymbol{\theta})$, usualmente o conhecimento proveniente da experiência subjetiva do projetista é o fator preponderante para a escolha da distribuição *a priori*. Técnicas para auxiliar o projetista nessa tarefa incluem o método do histograma, o método da função de distribuição e o método de verossimilhança relativa [4]. Embora úteis, essas técnicas costumam gerar distribuições complicadas e de difícil tratamento analítico, o que prejudica o procedimento posterior de otimização.

Prioris conjugadas

Para assegurar a tratabilidade da distribuição *a posteriori*, uma estratégia popular é adotar prioris com a mesma estrutura algébrica da verossimilhança, garantindo que o produto entre ambas, do qual depende a posteriori, tenha uma forma de simples manipulação. Por exemplo, se a verossimilhança é gaussiana, a escolha de uma priori também gaussiana gera uma posteriori gaussiana, que é uma distribuição com boas propriedades, de fácil otimização ou amostragem.

Voltemos ao modelo linear generalizado da Seção 1.3. Como o ruído é gaussiano, também o é a verossimilhança. Assumindo um priori para $\boldsymbol{\theta}$ gaussiana com média $\mathbf{m}_\theta$ e matriz de covariância $\mathbf{C}_\theta$, a posteriori seria então o produto de gaussianas:

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p_v(\mathbf{x} - \mathbf{G}\boldsymbol{\theta})p(\boldsymbol{\theta}) = \mathrm{N}(\mathbf{x}|\mathbf{G}\boldsymbol{\theta}, \sigma_v^2\mathbf{I})\mathrm{N}(\boldsymbol{\theta}|\mathbf{m}_\theta, \mathbf{C}_\theta), \tag{1.16}$$

cujo resultado é também uma gaussiana com parâmetros modificados:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \mathrm{N}\left(\boldsymbol{\theta}\,\middle|\,\mathbf{G}^T\mathbf{G} + \sigma_v^2\mathbf{C}_{\boldsymbol{\theta}}^{-1})^{-1}(\mathbf{G}^T\mathbf{x} + \sigma_v^2\mathbf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{m}_{\boldsymbol{\theta}}, \frac{\mathbf{G}^T\mathbf{G}}{\sigma_v^2} + \mathbf{C}_\theta^{-1}\right) \tag{1.17}$$

Caso a variância do ruído σ_v^2 nesse modelo seja desconhecida, podemos descrevê-la estatisticamente. Olhando para expressão da verossimilhança, vemos que σ_v^2 aparece de forma similar a uma distribuição gama inversa, uma que tem suporte no eixo real positivo e definida por parâmetros α e β:

$$p(x|\alpha, \beta) = \mathrm{IG}(x|\alpha, \beta) = \frac{\beta^\alpha}{\tau(\alpha)}x^{-\alpha-1}\exp\left(\frac{-\beta}{x}\right) \tag{1.18}$$

Sendo assim, a priori conjugada para σ_v^2 é uma gama-inversa $p(\sigma_v^2|\alpha_v, \beta_v)$. Com essa priori, a posteriori

será também uma gama-inversa com parâmetros modificados:

$$p(\sigma_v^2|\mathbf{y},\boldsymbol{\theta}) = \text{IG}\left(\sigma_v^2 \middle| \alpha_v + \frac{N}{2}, \beta_v + \frac{(\mathbf{y}-\mathbf{G}\boldsymbol{\theta})^T(\mathbf{y}-\mathbf{G}\boldsymbol{\theta})}{2}\right), \tag{1.19}$$

onde N é o número de elementos em $\mathbf{x}$.

Priori não-informativa

Como sugere o nome, priori não-informativa é uma distribuição que idealmente não contém qualquer informação sobre os possíveis valores dos parâmetros. A escolha aparentemente mais óbvia para uma priori não-informativa seria uma distribuição uniforme associando igual probabilidade a todos os valores no espaço amostral do parâmetro, o que produziria uma posteriori com formato idêntico à verossimilhança. Razoável para parâmetros discretos, essa escolha tende a ser problemática em espaços contínuos, uma vez que a distribuição uniforme não é invariante a transformações bijetivas de variáveis. Em outras palavras, se $\boldsymbol{\theta}$ é uniforme, a distribuição de $\boldsymbol{\phi} = f(\boldsymbol{\theta})$ não será mais uniforme [4]. Esse resultado é indesejável porque a ausência de conhecimento sobre $\boldsymbol{\theta}$ deveria implicar total desconhecimento sobre qualquer variável obtida pela transformação de $\boldsymbol{\theta}$.

Para eliminar esse inconveniente com a escolha da uniforme como priori, Jeffreys introduziu uma classe de distribuições que é ao mesmo tempo bastante vaga e invariante a transformações um-para-um [4]. A priori de Jeffreys associada à variável $\boldsymbol{\theta}$ é dada por:

$$p(\boldsymbol{\theta}) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2}, \tag{1.20}$$

onde $\mathbf{I}(\boldsymbol{\theta})$ é a matriz informação de Fisher [1] de $\boldsymbol{\theta}$, definida por:

$$I(\boldsymbol{\theta}) = E_{\mathbf{X}|\boldsymbol{\theta}}\left[-\frac{\partial^2 \ln\left(p(\mathbf{X}|\boldsymbol{\theta})\right)}{\partial \boldsymbol{\theta}^2}\right], \tag{1.21}$$

onde $E_{\mathbf{X}|\boldsymbol{\theta}}$ denota o valor esperado da variável $\mathbf{X}|\boldsymbol{\theta}$. Note que a priori de Jeffreys depende apenas da verossimilhança. No caso de modelos invariantes com relação à escala, como o desvio-padrão da Gaussiana no modelo linear generalizado, obtém-se a seguinte *priori* de Jeffreys:

$$p(\sigma_v) \propto \frac{1}{\sigma_v} \tag{1.22}$$

Há outras alternativas para a escolha de prioris não-informativas. Jaynes [5] propôs que o critério de máxima entropia fosse usado na especificação da distribuição *a priori*. Bernardo [6] também defende que a escolha da priori não-informativa deva ser feita através de critérios baseados na teoria da informação. Na maioria dos casos, a distribuição gerada é imprópria, isto é, sua integral é infinita, como na Equação (1.22).

Esse fato nem sempre é um inconveniente, visto que estamos interessados, muitas vezes, na forma da distribuição, e não nos seus valores exatos.

1.5 Simulação Estatística através de Monte Carlo por Cadeias de Markov (MCMC)

Embora seja possível resolver problemas práticos apenas com os conceitos vistos aqui, em boa parte das aplicações as distribuições envolvidas são multivariáveis e multi-modais para as quais as técnicas clássicas de otimização são inadequadas. Não raro a modelagem realista de sistemas físicos exige o uso de modelos hierárquicos de vários níveis, que produzem distribuições com distribuições de formato complicado e intratáveis analiticamente. Situações assim requerem técnicas numéricas mais sofisticadas, dentre as quais destacamos o algoritmo EM (*Expectation-Maximization*), o amostrador de Gibbs e os algoritmos de Metropolis-Hastings (MH). Esses dois últimos baseiam-se em Monte Carlo via Cadeias de Markov (*Markov-Chain Monte Carlo*, MCMC) [7], uma classe de métodos que consiste no projeto de uma cadeia de Markov para obtenção de amostras de uma certa distribuição alvo, as quais são posteriormente usadas para obtenção de informações de interesse via técnicas de Monte Carlo.

1.6 Técnicas de Monte Carlo

De forma geral, técnicas de Monte Carlo consistem em gerar amostras i.i.d. (independentes e identicamente distribuídas) de uma certa distribuição de interesse, para que possam ser usadas para obter uma aproximação de alguma característica da distribuição. De posse de um conjunto de amostras $X = \{x^{(1)}, \ldots, x^{(N)}\}$ de uma certa distribuição $p(x)$, podemos aproximar uma integral do tipo

$$I(f) = \int f(x)p(x)dx \tag{1.23}$$

por um um somatório:

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}). \tag{1.24}$$

É possível mostrar que o estimador acima é não enviesado e, pela lei dos grandes números, converge para a integral da Equação (1.23) *quase certamente* (isto é, com probabilidade 1) quando N tende a infinito.

Pela Eq. 1.23 vemos que Monte Carlo pode ser usado para cálculo numérico de integrais de um tipo qualquer, bastante para isso escrever o integrando como o produto de uma função $f(x)$ com outra uma positiva e de integral unitária $p(x)$. Se quisermos por exemplo usar Monte Carlo para encontrar uma aproximação para a área de um círculo de centro em $(0, 0)$ e raio 1, podemos considerar $p(x)$

uma distribuição uniforme no quadrado de vértices $(1,1)$, $(1,-1)$, $(-1,1)$ e $(-1,-1)$ e $f(x)$ uma função indicadora que retorna 1 se x estiver no interior do círculo e 0 caso contrário. A geração de amostras de $p(x)$ é feita trivialmente com funções padronizadas existente nas linguagens de programação mais usuais. A avaliação de $f(x)$ para um x qualquer também é simples: basta fazer $f(x) = 1$ se $x_1^2 + x_2^2 > 1$ e $f(x) = 0$ caso contrário. Gerando $N = 1000$ amostras de $p(x)$ e calculando o somatório da Eq. 1.24, encontramos $I_N(f) \approx 3.13988$, que aproxima-se notavelmente do valor correto da área: $\pi r^2 \approx 3.14159265$.

1.7 Cadeias de Markov

Esse último exemplo é um caso fortuito em que a amostragem de $p(x)$ foi simples – distribuição uniforme. As técnicas envolvendo cadeias de Markov permitem realizar a tarefa de amostragem em casos mais gerais e difíceis, de forma indireta. Começamos considerando cadeias de Markov com espaço de estados discreto e em seguida estenderemos os resultados para o caso contínuo.

1.7.1 Conceitos

Uma Cadeia de Markov é um processo aleatório discreto no tempo que apresenta em que o estado atual da cadeia depende unicamente do estado imediatamente anterior – condição conhecida como *propriedade de Markov*. Mais formalmente, seja $X^{(n)}$ a variável aleatória que representa o estado da cadeia no instante n e seja S o espaço de estados para as variáveis $X^{(n)}$. Então:

$$P(X^{(n)} \in A^{(n)} | X^{(n-1)} \in A^{(n-1)}, \ldots, X^{(0)} \in A^{(0)}) = P(X^{(n)} \in A^{(n)} | X^{(n-1)} \in A^{(n-1)}), \tag{1.25}$$

para quaisquer $A^{(0)}, \ldots, A^{(n)} \in S$.

Se S é um conjunto contável a cadeia de Markov é dita discreta. Consideremos inicialmente um espaço de estados $S = \{s_1, \ldots, s_N\}$. As probabilidades de transição de um estado da cadeia para outro, no instante n, definem a *matriz de transição* $\mathbf{T}_n$, cujo elemento da linha i e da coluna j é dado por:

$$T_n(i|j) = P(X^{(n)} = s_i | X^{(n-1)} = s_j). \tag{1.26}$$

Como cada coluna dessa matriz contém a probabilidade associada a cada um dos elementos do espaço amostral, a soma dos elementos em cada coluna é igual a 1. Matrizes com essa propriedade são chamadas matrizes estocásticas [8] e possuem uma série de propriedades importantes para a análise de algoritmos baseados em MCMC. Uma delas é a existência de pelo menos um autovalor igual a 1. Além disso, no caso em que todos os elementos de $\mathbf{T}_n$ são maiores que zero – caso em que é sempre possível passar de um estado qualquer para outro qualquer –, pode-se mostrar que todos os demais autovalores são distintos e menores que 1.

A distribuição de probabilidade no instante n, denotada por $P_n(i)$, definida para $i = \{1, \ldots, N\}$, pode ser obtida a partir da distribuição de probabilidade no instante $(n-1)$, através de:

$$P_n(i) = \sum_{j=1}^{N} T_n(i|j) P_{n-1}(j), \tag{1.27}$$

que pode ser escrita vetorialmente como:

$$\mathbf{P}_n = \mathbf{T}_n \mathbf{P}_{n-1}, \tag{1.28}$$

em que $\mathbf{P}_n = [P_n(1) \ldots P_n(N)]^T$.

Se $\mathbf{T}_n$ é independente de n, dizemos que a cadeia de Markov é *homogênea* e a matriz de transição passa a ser denotada simplesmente por $\mathbf{T}$. Nesse caso, a aplicação da Equação (1.28) n vezes fornece:

$$\mathbf{P}_n = \mathbf{T}^n \mathbf{P}_0, \tag{1.29}$$

em que $\mathbf{P}_0$ é a distribuição do estado inicial da cadeia.

Para o desenvolvimento de algoritmos MCMC, a Cadeia de Markov deve possuir as duas propriedades abaixo:

- Irredutibilidade: Partindo de qualquer estado, existir uma probabilidade não-nula de a cadeia mover-se para qualquer outro estado em um número finito de passos. Isso equivale a termos $T^n(i|j) > 0$, para algum n, de forma que a probabilidade de transição em n passos do estado s_i para o estado s_j seja diferente de zero.
- Aperiodicidade: A cadeia não ficar presa em ciclos.

1.7.2 Distribuição invariante

Uma distribuição é dita invariante (ou estacionária) se permanece fixa sob a aplicação da matriz de transição. No desenvolvimento de algoritmos MCMC, estamos interessados em construir cadeias de Markov que façam com que uma dada distribuição seja invariante. Denotando a distribuição invariante por $\pi(i)$, devemos ter:

$$\pi(i) = \sum_{j=1}^{N} T(i|j) \pi(j). \tag{1.30}$$

Vetorialmente, podemos escrever $\boldsymbol{\pi} = \mathbf{T}\boldsymbol{\pi}$, donde vemos que $\boldsymbol{\pi}$ é um autovetor associado ao autovalor $\lambda = 1$. Para se determinar $\boldsymbol{\pi}$ unicamente, impõe-se a condição de que a soma de seus elementos seja igual a 1.

Para algoritmos MCMC, é usual impor a condição *detailed balance* para assegurar que uma dada distribuição seja invariante. Essa condição estabelece que a probabilidade de a cadeia passar de um

estado s_i no instante $(n-1)$ para o estado s_j no instante (n) é igual à probabilidade de a transição inversa ocorrer, isto é:

$$\pi(j)T(i|j) = \pi(i)T(j|i). \tag{1.31}$$

Para ver que $\pi(i)$ é distribuição invariante, basta somar os dois membros da igualdade acima para todos os valores possíveis de j, e perceber que o resultado é a Equação (1.30). Embora mais restritiva que a Equação (1.30), essa condição é mais simples de se impor para algoritmos MCMC.

1.7.3 Ergodicidade

Além de garantir que $\pi(i)$ seja a distribuição invariante desejada, devemos assegurar que $P_n(i)$ convirja para $\pi(i)$ quando n tender a infinito, qualquer que seja a distribuição inicial $P_0(i)$. Nesse caso, dizemos que $\pi(i)$ é a distribuição-limite da cadeia.

Essa propriedade é conhecida como *ergodicidade*. Para uma cadeia de Markov ser ergódica, é necessário que a cadeia seja aperiódica e irredutível. No caso discreto, basta termos os autovalores de $\mathbf{T}$ todos distintos, podendo assim escrever a distribuição inicial usando os autovetores de $\mathbf{T}$ como base e em seguida calcular a expressão $\mathbf{T}^n\mathbf{P}_0$ [9]:

$$\mathbf{P}_0 = \boldsymbol{\pi} + c_2\mathbf{v}_2 + \ldots + c_N\mathbf{v}_N; \tag{1.32}$$

$$\mathbf{P}_n = \mathbf{T}^n\mathbf{P}_0 = \boldsymbol{\pi} + c_2\lambda_2^n\mathbf{v}_2 + \ldots + c_N\lambda_N^n\mathbf{v}_N. \tag{1.33}$$

Na primeira equação usamos o fato de que $\boldsymbol{\pi}$ é o autovetor associado ao autovalor 1. Como sabemos que todos os demais autovalores são menores que 1, concluímos que:

$$\lim_{n\to\infty}\mathbf{P}_n = \lim_{n\to\infty}\mathbf{T}^n\mathbf{P}_0 = \boldsymbol{\pi}. \tag{1.34}$$

1.7.4 Cadeias de Markov para espaço contínuo de estados

Para o caso em que o espaço de estados S é contínuo, as propriedades descritas nas seções anteriores são expressas através de densidades de probabilidade. A propriedade de Markov é definida por:

$$p(x^{(n)}|x^{(n-1)},\ldots,x^{(0)}) = p(x^{(n)}|x^{(n-1)}). \tag{1.35}$$

O *núcleo de transição* $K_n(x|y)$ no instante n é definido por:

$$K_n(x|y) = p_{X^{(n)}}(x|X^{(n-1)} = y), \tag{1.36}$$

em que $p_{X^{(n)}}(x)$ denota a densidade de probabilidade da variável aleatória $X^{(n)}$.

Assim, a distribuição do estado da cadeia no instante n é dada por:

$$p_n(x) = \int_{y \in S} K_n(x|y) p_{n-1}(y) dy. \tag{1.37}$$

No caso de a cadeia ser homogênea, K_n independe de n e a condição *detailed balance* se torna:

$$\int_A \int_B K(x|y)\pi(y)dydx = \int_B \int_A K(y|x)\pi(x)dxdy, \tag{1.38}$$

para quaisquer conjuntos A e B pertencentes a S.

Como no caso discreto, a condição *detailed balance* é suficiente para que $\pi(i)$ seja uma distribuição invariante da cadeia de Markov definida por $K(i|j)$. Para garantir que a distribuição invariante seja também a distribuição-limite, a cadeia deve ser aperiódica e irredutível.

1.8 Amostrador de Gibbs

O amostrador de Gibbs (*Gibbs Sampler*) [10] é um caso especial de MCMC, indicado para os casos em que a distribuição conjunta é mais difícil de amostrar do que as condicionais. A técnica consiste em particionar a variável conjunta em diversos componentes (possivelmente multivariáveis) e obter amostras das distribuições condicionais de cada componente, considerando os demais fixos. O processo é repetido usando os últimos valores amostrados de cada componente como condicionantes da distribuição dos demais componentes.

Seja $\pi(\boldsymbol{\theta})$ a distribuição conjunta da qual se deseja obter amostras. A variável $\boldsymbol{\theta}$ é então particionada em k componentes, de tal forma que $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$. A i-ésima iteração do amostrador de Gibbs pode ser expressa como:

$$\boldsymbol{\theta}_1^{(i)} \sim \pi(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(i-1)}, \ldots, \boldsymbol{\theta}_k^{(i-1)}) \tag{1.39}$$

$$\boldsymbol{\theta}_2^{(i)} \sim \pi(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_k^{(i-1)}) \tag{1.40}$$

$$\vdots$$

$$\boldsymbol{\theta}_k^{(i)} \sim \pi(\boldsymbol{\theta}_k | \boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{k-1}^{(i)}), \tag{1.41}$$

em que o símbolo $\sim$ indica que a variável da esquerda é uma amostra da distribuição à direita.

Para ver que $\pi(\boldsymbol{\theta})$ é uma distribuição invariante em cada operação acima, vamos calcular a distribuição de $\boldsymbol{\theta}$ após a primeira operação, definida na Equação (1.39). Supondo que a distribuição da cadeia de

Markov no final da iteração $(i-1)$ é $\pi(\boldsymbol{\theta})$, isto é: $p(\boldsymbol{\theta}^{(i-1)}) = \pi(\boldsymbol{\theta}^{(i-1)})$, obtemos:

$$\begin{aligned} p(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i-1)}, \ldots, \boldsymbol{\theta}_k^{(i-1)}) &= p(\boldsymbol{\theta}_1^{(i)} | \boldsymbol{\theta}_2^{(i-1)}, \ldots, \boldsymbol{\theta}_k^{(i-1)}) p(\boldsymbol{\theta}_2^{(i-1)}, \ldots, \boldsymbol{\theta}_k^{(i-1)}) \\ &= \pi(\boldsymbol{\theta}_1^{(i)} | \boldsymbol{\theta}_2^{(i-1)}, \ldots, \boldsymbol{\theta}_k^{(i-1)}) \pi(\boldsymbol{\theta}_2^{(i-1)}, \ldots, \boldsymbol{\theta}_k^{(i-1)}) \\ &= \pi(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i-1)}, \ldots, \boldsymbol{\theta}_k^{(i-1)}). \end{aligned} \tag{1.42}$$

Logo, a transição gerada pelo amostrador de Gibbs preserva a distribuição de $\boldsymbol{\theta}$. O mesmo raciocínio aplicado às operações seguintes indica que, a cada amostragem, a distribuição resultante permanece igual a $\pi(\boldsymbol{\theta})$. Sob condições pouco restritivas, mostra-se que a cadeia é ergódica, isto é, após a convergência, as amostras geradas em cada iteração correspondem a amostras da distribuição conjunta $\pi(\boldsymbol{\theta})$.

1.9 Algoritmo de Metropolis-Hastings

Nem sempre as distribuições condicionais necessárias para o amostrador de Gibbs são fáceis de se obter. Nesses casos, o algoritmo de Metropolis-Hastings (MH), proposto por Nicholas Metropolis em 1953 [11] e generalizado por Hastings em 1970 [12], é mais indicado. A ideia do algoritmo é inicialmente obter amostras de uma distribuição auxiliar supostamente mais simples que a distribuição de interesse, e em seguida decidir aceitar ou rejeitar essa amostra dependendo de um critério probabilístico, a ser especificado.

Pelo algoritmo MH, uma amostra x^* é obtida a partir de uma distribuição proposta, designada por $q(x^*|x^{(i)})$, em que $x^{(i)}$ é o estado atual da cadeia de Markov. Essa amostra é aceita com uma probabilidade α dada por:

$$\alpha(x^{(i)}, x^*) = \min\left(1, \frac{\pi(x^*) q(x^{(i)}|x^*)}{\pi(x^{(i)}) q(x^*|x^{(i)})}\right). \tag{1.43}$$

Se a amostra gerada for aceita, o novo estado da cadeia é $x^{i+1} = x^*$; em caso contrário, a cadeia permanece no seu estado atual, isto é, $x^{(i+1)} = x^{(i)}$. O núcleo de transição é dado por:

$$K(x^{(i+1)}|x^{(i)}) = q(x^{(i+1)}|x^{(i)}) \alpha(x^{(i)}, x^{(i+1)}) + \delta_{x^{(i)}}(x^{(i+1)}) r(x^{(i)}), \tag{1.44}$$

em que

$$r(x^{(i)}) = \int_{x^* \in S} q(x^*|x^{(i)}) \left(1 - \alpha(x^{(i)}, x^*)\right) dx^* \tag{1.45}$$

é a probabilidade de a cadeia permanecer no estado atual. A expressão $\delta_{x^{(i)}}(x^{(i+1)})$ é a função impulso unitário (delta de Dirac) localizada em $x^{(i)}$ e aplicada em $x^{(i+1)}$, que indica uma distribuição "concentrada" em $x^{(i)}$.

Pode-se mostrar que esse núcleo de transição satisfaz a condição *detailed balance*, e portanto $\pi(x)$ é

uma distribuição invariante da cadeia. Para garantir que $\pi(x)$ seja também a distribuição-limite, temos que verificar a irredutibilidade e aperiodicidade da cadeia. Como o algoritmo sempre permite a rejeição, segue que a cadeia é aperiódica. Para assegurar irredutibilidade, é preciso que o suporte de $q(\cdot)$ inclua o suporte de $\pi(\cdot)$ [13].

A eficiência do algoritmo MH depende fundamentalmente da escolha da proposta, $q(\cdot)$. É usual escolher como proposta uma gaussiana centrada no estado atual, isto é, $q(x^*|x^{(i)}) = N(x^*|x^{(i)}, \sigma_q^2 I)$. A escolha da variância σ_q^2 é crucial. Se $q(\cdot)$ é muito estreita, apenas estados próximos ao máximo de $\pi(x)$ são visitados. Por outro lado, se $q(\cdot)$ é muito ampla, o percentual de amostras rejeitadas é muito alto e, consequentemente, as amostras geradas serão altamente correlacionadas entre si, invalidando a hipótese de independência das amostras, necessária para a estimação de Monte Carlo. Para resolver esse problema, recomenda-se descartar uma fração das amostras, reduzindo assim a correlação entre elas, mas, por outro lado, aumentando o tempo para convergência.

1.10 Análise de convergência

Os algoritmos apresentados até agora garantem que os estados da cadeia de Markov sejam amostras da distribuição-limite quando o número de iterações tende a infinito. Como na prática precisamos usar um número finito de iterações, é importante saber quando a cadeia está suficientemente próxima da distribuição limite para que suas amostras possam ser usadas para estimação de Monte Carlo.

Muitos trabalhos tem sido desenvolvidos em torno desse problema. Métodos teóricos para diagnóstico de convergência têm sido propostos, mas os resultados tiveram até o momento pouco impacto prático [8]. Na maioria das aplicações, usam-se mais frequentemente métodos informais de convergência, que consistem de análise estatística dos dados gerados pela cadeia. Embora simples, essas técnicas não permitem garantir a convergência de forma geral.

Dentre os métodos informais, três abordagens podem ser usadas. A primeira baseia-se na realização de n cadeias paralelas. Calcula-se o histograma de n amostras na m-ésima iteração e se o compara com o histograma obtido k iterações adiante. Se os histogramas forem bastante similares, considera-se que a convergência foi atingida. O valor de k deve ser grande o bastante para evitar que a correlação entre os estados sucessivos da cadeia causem uma falsa impressão de similaridade entre os histogramas. O critério para comparação dos histogramas é flexível, sendo uma possibilidade a divergência de Kullback-Leibler.

Outro método baseia-se na construção de uma única cadeia e no cálculo da média ergódica das amostras obtidas. Espera-se que após a convergência, a média se aproxime de um certo valor constante. Portanto, através da análise visual do gráfico das médias, é possível diagnosticar a convergência.

A terceira técnica consiste na análise visual da evolução dos parâmetros da cadeia. Após a convergência, esses parâmetros devem exibir um padrão constante, que pode ser usado para diagnosticá-la.

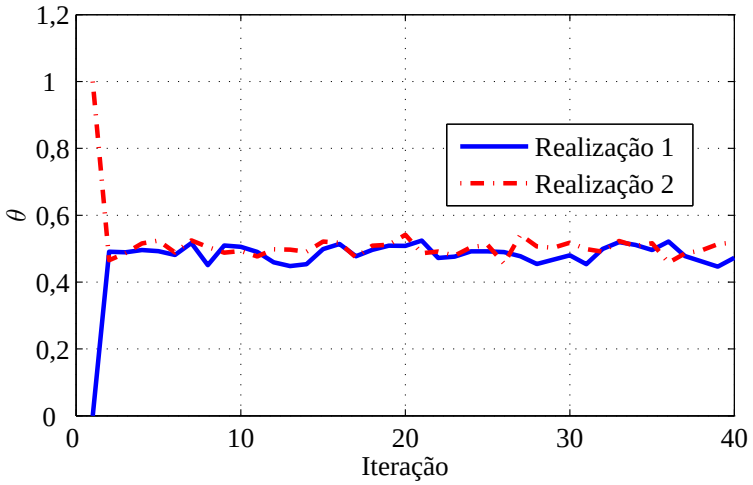

Figura 1.4: Evolução de um parâmetro de uma cadeia de Markov a partir de diferentes valores iniciais. Vemos que as amostras tendem a variar na mesma região a partir de certo ponto, independente da inicialização, evidenciando a convergência do algoritmo.

Analisando a Figura 1.4, que mostra a evolução de um parâmetro de uma cadeia de Markov partindo de pontos iniciais diferentes, podemos concluir que a convergência ocorre em torno da iteração 90.

1.11 Deteção de Símbolos Digitais em Canais Seletivos em Frequência

Consideremos agora como os conceitos visto até aqui podem nos auxiliar na solução do recorrente problema deteção de sinais em sistemas de telecomunicações digitais sujeito a um canal ruidoso e seletivo em frequência. Para nos concentrarmos nas ideias centrais evitando complicações demasiadas, vamos nos restringir à transmissão digital BPSK (*Binary Phase Shift-Keying*), na qual modulação de fase é usada para transmitir símbolos binários x_1, x_2, ..., x_N, com $x_k \in \{+1, -1\}$. Sendo o canal seletivo em frequência, o sinal recebido y_n é distorcido e pode ser escrito como combinação linear dos dados enviados deslocadas no tempo, adicionada ao ruído do canal:

$$y_n = \sum_{l=0}^{L-1} h_l x_{n-l} + v_n, \tag{1.46}$$

em que $\mathbf{h} = \{h_1, \ldots, h_L\}$ representa a resposta ao impulso do canal de comprimento L – aqui assumido invariante no tempo – e $\mathbf{v} = \{v_1, \ldots, v_N\}$ são amostras i.i.d de ruído branco gaussiano de média zero e variância σ_v^2.

Do ponto de vista Bayesiano, o problema consiste em estimar a sequência enviada $\mathbf{x} = \{x_1, \ldots, x_N\}$ com base no sinal recebido $\mathbf{y} = \{y_1, \ldots, y_N\}$, tratando como *nuisance* o conjunto de parâmetros $\boldsymbol{\theta} = \{h_0, \ldots, h_L, \sigma_v^2\}$. A solução do problema começa com a obtenção da distribuição *a posteriori* das quantias

desconhecidas – que no caso inclui os dados $\mathbf{x}$ e os parâmetros *nuisance* $\boldsymbol{\theta}$. Pelo teorema de Bayes temos que:

$$p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{1.47}$$

em que

$$p(\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{1.48}$$

e $p(\mathbf{x})$ e $p(\boldsymbol{\theta})$ são as distribuições *a priori* dos dados enviados e dos parâmetros *nuisance*, respectivamente. Sendo os símbolos +1 e -1 equiprováveis, a priori para $\mathbf{x}$ é constante para todas as possíveis combinações de símbolos enviados. As prioris para $\mathbf{h}$ e σ_v^2 dependerão do ambiente do canal em questão (rural ou urbano, por exemplo), da banda de transmissão, e do nível esperado de razão sinal-ruído.

Além das prioris, a especificação da posteriori depende do cálculo da verossimilhança. No caso, a situação é bem parecida com o modelo linear geral discutido na Seção 1.3 e a verossimilhança é dada por:

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = p_v\left(\mathbf{y} - \mathbf{H}\mathbf{x}\right), \tag{1.49}$$

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_v^2}(\mathbf{y} - \mathbf{H}\mathbf{x})^T(\mathbf{y} - \mathbf{H}\mathbf{x})\right\}. \tag{1.50}$$

Substituindo as prioris e a verossimilhança em 1.47, a posteriori é obtida. Essa é apenas a primeira parte do trabalho. O segundo passo é a integração dos parâmetros *nuisance* para a obtenção da distribuição condicional dos dados enviados sendo conhecidos os dados recebidos:

$$p(\mathbf{x}|\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \tag{1.51}$$

Mesmo nessa situação simplificada com o canal constante e ruído branco, a integral acima não pode ser resolvida analiticamente, e os métodos tradicionais de integração numérica são inadequadas devido à alta dimensão de $\boldsymbol{\theta}$. Outra dificuldade é a estimativa da sequência de dados mais prováveis. A solução trivial, uma busca exaustiva por todas as 2^N sequências possíveis, é claramente impraticável para valores típicos de N.

Técnicas de MCMC permitem superar essas duas dificuldades. Em particular, o amostrador de Gibbs realiza numericamente a integral acima de forma indireta, através da amostragem das variáveis em questão a partir de suas distribuições condicionais. O mesmo algoritmo permite reduzir dramaticamente a custo computacional na estimativa de $\mathbf{x}$ se for adotada a amostragem sequencial em blocos de tamanho reduzido.

As operações de mais alto nível do procedimento MCMC estão resumidas no Algoritmo 1, em que o símbolo $\sim$ denota que a variável à esquerda é uma amostra da distribuição à direita. Em seguida, detalhamos cada uma de suas operações do algoritmo.

Algorithm 1 Algoritmo MCMC para deteção de sinais em canais de comunicação seletivos em frequência.

1: Inicialização $\mathbf{x}^{(0)}$, $\mathbf{h}^{(0)}$, $\sigma_v^{2(0)}$;
2: **for** $i = 1$ to I **do**
3: $\quad \mathbf{x}^{(i+1)} \sim p(\mathbf{x}|\mathbf{y}, \mathbf{h}^{(i)}, \sigma_v^{2(i)})$
4: $\quad \mathbf{h}^{(i+1)} \sim p(\mathbf{h}|\mathbf{y}, \mathbf{x}^{(i+1)}, \sigma_v^{2(i)})$
5: $\quad \sigma_v^{2(i+1)} \sim p(\sigma_v^2|\mathbf{y}, \mathbf{x}^{(i+1)}, \mathbf{h}^{(i+1)})$
6: **end for**

1.11.1 Amostragem de x

A amostragem direta de $\mathbf{x}$ seria impraticável por exigir o cálculo de 2^N valores associados a cada sequência possível. Uma maneira de reduzir esse custo é explorar a flexibilidade do amostrador de Gibbs e realizar essa operação em sub-blocos de tamanho reduzido. A ideia é particionar a sequência de N elementos em b blocos de tamanho B de tal forma que $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_B\}$. Em seguida fazemos a amostragem de cada sub-bloco $\mathbf{x}_j$ sequencialmente. Na amostragem do sub-bloco $\mathbf{x}_j$, os demais blocos são considerados conhecidos e iguais aos previamente amostrados. Como cada bloco tem B elementos, agora a amostragem requer o cálculo das probabilidades condicionais para as 2^B subsequências possíveis dentro de um bloco.

Amostragem de h

Para calcularmos a distribuição condicional de $\mathbf{h}$ é conveniente escrever a Eq. 1.46 de uma forma alternativa explicitando a dependência dos dados com o parâmetro $\mathbf{h}$:

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{v} \tag{1.52}$$

em que $\mathbf{X}$ é formada por elementos de $\mathbf{x}$. Assim, a condicional total de $\mathbf{h}$ é dada por:

$$p(\mathbf{h}|\mathbf{y}, \mathbf{x}, \sigma_v^2) = p_v(\mathbf{y} - \mathbf{X}\mathbf{h})p(\mathbf{h}) = \mathrm{N}(\mathbf{h}|\mathbf{m}_h, \mathbf{C}_h), \tag{1.53}$$

onde $\mathbf{m}_h = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ e $\mathbf{C}_h = \sigma_v^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Amostragem de σ_v^2

Na discussão sobre eliminação de parâmetros (Seção 1.4.2) vimos que uma priori do tipo gama inversa para a variância σ_v^2 gera uma posteriori do mesmo formato com parâmetros modificados. Como no momento da amostragem de σ_v^2, $\mathbf{h}$ e $\mathbf{x}$ são conhecidos, podemos usar a fórmula da Eq. 1.19 para calcular a distribuição condicional de σ_v^2, apenas substituindo $\mathbf{G}$ por $\mathbf{X}$, $\boldsymbol{\theta}$ por $\mathbf{h}$ e $\mathbf{x}$ por $\mathbf{y}$.

Simulações

Vamos ilustrar o algoritmo MCMC numa situação de canal severamente seletivo em frequência e razão sinal-ruído (SNR) variando numa ampla faixa. Escolhemos o canal de comprimento $L = 11$ com resposta ao impulso definida na Fig. 1.5 e variância do ruído $\sigma_v^2 = 0.1$ correspondendo a uma SNR de 10 dB.

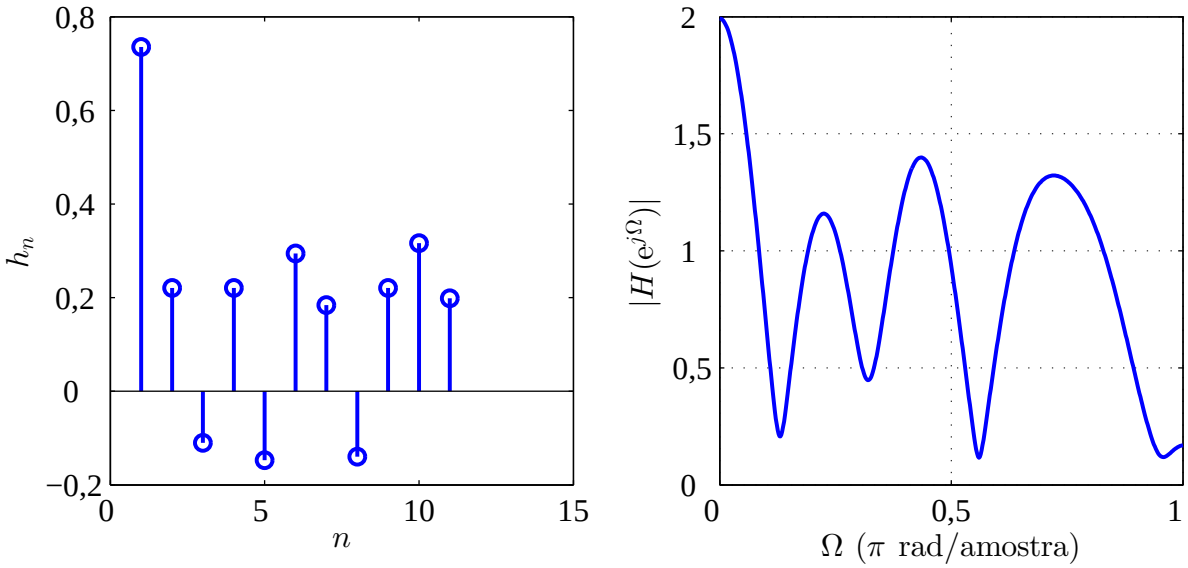

Figura 1.5: Caracterização dos multi-percursos no canal de comunicação. (a) Resposta ao impulso do canal. (b) Módulo da resposta em frequência do canal.

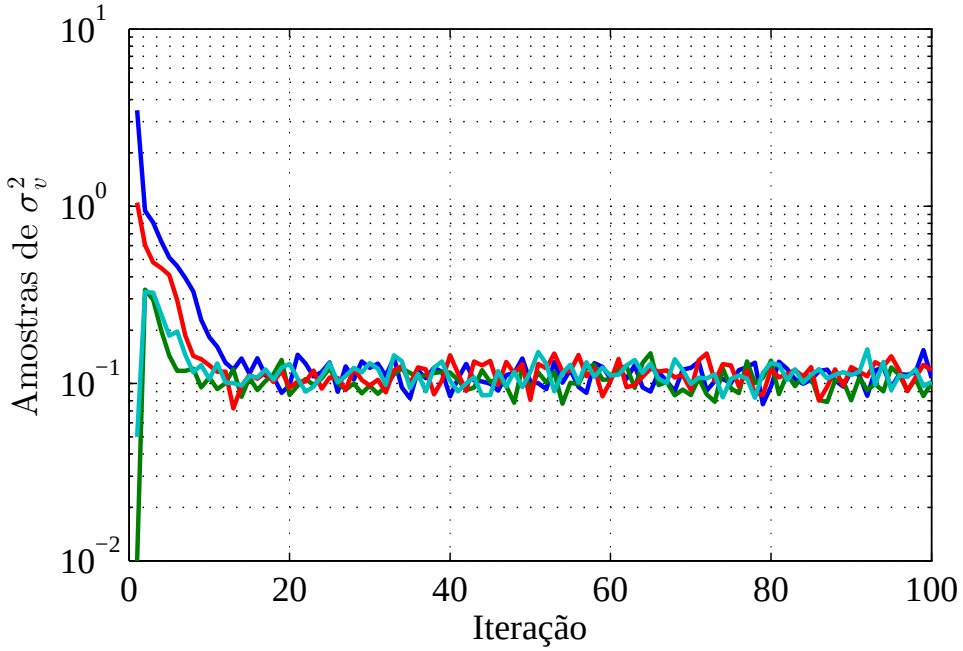

Figura 1.6: Evolução da variância do ruído σ_v^2 através do amostrador de Gibbs, para diversas inicializações. A cadeia converge após poucas dezenas de iterações.

Inicializamos os parâmetros com $\mathbf{x}^0 = \text{sign}(\mathbf{y})$, $\sigma_v^{2(0)} = \text{IG}(\sigma_v^2|2,1)$, correspondente à sua priori com

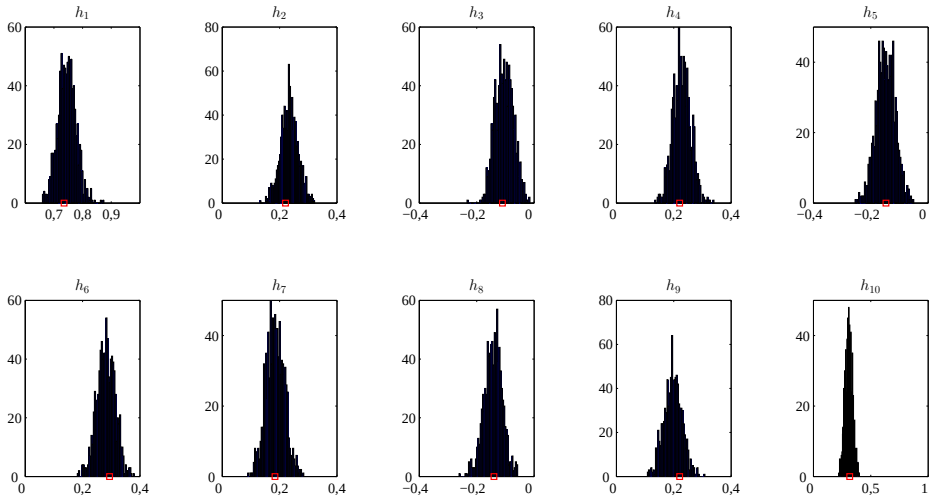

Figura 1.7: Histogramas de cada elemento dos coeficientes do modelo do canal, gerados a partir pelo amostrador de Gibbs. As amostras concentram-se em torno do valor correto (em vermelho).

$\alpha_v = 2$ e $\beta = 1$. O algoritmo é executado por 1000 iterações partindo de vários valores iniciais para σ_v^2. A evolução de σ_v^2 ao longo do tempo para as diversas inicializações (Fig. 1.6) nos permite concluir que a cadeia em geral converge após algumas dezenas de iterações. Avaliando as amostras geradas após um *burn-in* de 100 iterações, concluímos que o algoritmo estima satisfatoriamente os parâmetros do modelo, como pode ser comprovado pelos histogramas da Fig. 1.7. A partir dos valores amostrados para $\mathbf{x}$ podemos facilmente calcular a estimativa de máximo *a posteriori* para cada símbolo enviado. Nesse exemplo, a taxa de erro de símbolo foi de 0.0977 %. Em comparação, um detetor simples que apenas avalia o sinal (positivo ou negativo) de $\mathbf{y}$ para estimar $\mathbf{x}$ produziria uma taxa de erro de 15.03 %.

1.12 Conclusão

Neste capítulo vimos os fundamentos de inferência Bayesiana e das técnicas de simulação estatística, com enfoque no problema de deteção de sinais em sistemas de comunicações. Escolhemos como estudo de caso um canal invariante no tempo com ruído gaussiano aditivo. Embora simplificado, esse modelo nos permite entender os principais aspectos da análise Bayesiana, a saber, a escolha das distribuições *a priori* para os parâmetros; o cálculo da distribuição *a posteriori* pelo teorema de Bayes, e a solução numérica do problema estimação através do amostrador de Gibbs. Ao contrário de técnicas tradicionais que evitam a modelagem estatística, o algoritmo resultante atinge estimativas ótimas, ao indiretamente minimizar a probabilidade de erro de transmissão. Além disso, em relação aos tradicionais equalizadores adaptativos, técnicas bayesianas tem a vantagem de produzirem estimadores cegos, pois não exigem a transmissão de uma sequência piloto treinamento – embora possa se beneficiar dela, se disponível. A desvantagem dessa

abordagem está no maior custo computacional requerido, em virtude das inversões de matrizes a cada iteração e do número de iterações para convergência. Esse último fator, contudo, pode ser largamente aliviado se adotarmos uma inicialização favorável – por exemplo através de uma técnica de estimação mais simples aplicada previamente.

Métodos bayesianos associados a técnicas de Markov Chain Monte Carlo tendem a se tornar cada vez mais ubíquos à medida que aumenta o poder computacional dos processadores. Ao permitir a modelagem completa e realista de um problema, tais métodos são atraentes em diversas aplicações em engenharia por oferecer estimativas de incomparável acurácia. Em particular, sistemas de comunicação que dependem crucialmente da confiabilidade na transmissão de dados, tendem a se beneficiar desse paradigma.

Referências

[1] L. Wasserman, *All of Statistics – A Concise Course in Statistical Inference.* Springer, 2010.

[2] S. Mcgrayne, *The Theory that Would not Die – How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines and Emerged Triumphant from Two Centuries of Controversy.* Yale University Press, 2011.

[3] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory.* New Jersey, USA: Prentice Hall, 1993.

[4] H. S. Migon and D. Gamerman, *Statistical Inference: An Integrated Approach.* Londres, Reino Unido: Hodder Arnold, 1999.

[5] E. Jaynes, "Prior probabilities," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 3, pp. 227–241, Set. 1978.

[6] J. M. Bernardo, "Reference posterior distributions for bayesian inference," *Journal of the Royal Statistical Society*, vol. 41, no. 2, pp. 113–147, Dez. 1979.

[7] C. Robert and G. Casella, *Monte Carlo Statistical Methods.* Nova Iorque, EUA: Springer, 2004.

[8] D. Gamerman and H. Lopes, *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference.* Boca Raton, USA: Chapman & Hall/CRC, 2006.

[9] R. Neal, "Probabilistic inference using markov chain monte carlo methods," University of Toronto, Department of Computer Science, Toronto, Canadá, Technical Report CRG-TR-93-1, 1993.

[10] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions and the bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[11] N. Metropolis, A. W. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[12] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[13] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, Jan.-Fev. 2003.