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

Flávio R. Ávila e Michel P. Tcheou

PROSAICO - DETEL/PEL - UERJ

1 de setembro de 2015

1 – Motivação

Problema 1: Qual a probabilidade de uma mulher ter câncer de mama sabendo que sua mamografia resultou anormal?
Dados

1. Prob. teste positivo na ausência de doença: 10 %
2. Prob. teste positivo com doença presente: 80 %
3. Prob. de uma mulher qualquer ter a doença: 0.4 %

Opções

(a) 1 %
(b) 3 %
(c) 50 %
(d) 80 %

Problema 2: Há três caixas, uma com duas bolas brancas, outra com duas bolas pretas e uma terceira com uma bola branca e outra preta. Escolhe-se aleatoriamente uma caixa e retira-se, também aleatoriamente, uma bola. Verifica-se que a bola é branca. Qual a probabilidade de que a bola remanescente naquela caixa também seja branca?

Opções

(a) 3/4
(b) 1/4
(c) 1/2
(d) 2/3

2 – Introdução

Inferência estatística

- Avaliar probabilidade de eventos incertos (passados ou futuros)
- Análise de riscos
- Predições
- Teste de hipóteses
- Regressão

2.1 – Escolas

- Frequentista
  - Probabilidade: frequência relativa
  - Parâmetros: constantes desconhecidas
- Bayesiana
  - Probabilidade: subjetiva ou pessoal
  - Parâmetros: variáveis aleatórias

Exemplo: Lançamento de um dado

Um dado é lançado e observa-se o número voltado para cima. Deseja-se obter as probabilidade de que esse valor seja primo, isto é, 2, 3 ou 5.

Solução Clássica

Princípio da razão insuficiente

"The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible."

— Pierre Laplace

No exemplo em questão

- Seis possíveis resultados igualmente prováveis
- Soma das probabilidades unitária
- Probabilidade de cada face: $1/6$
- Probabilidade de valor ser primo: razão entre o número de eventos favoráveis e a quantidade total eventos possíveis
- $p = 1/2$

Solução Frequentista

Frequência relativa de sucessos após longa sequência de lançamentos

$$p = \lim_{n \to \infty} \frac{\text{número de sucessos}}{n}$$

No exemplo em questão

- Conta-se o número de resultados primos numa longa e sequência de lançamentos
- Divide-se pelo número total de lançamentos
- Lei dos grandes números: $p \approx 1/2$

Solução Bayesiana

Probabilidade é o grau de confiança subjetiva sobre evento

Pode variar de pessoa pra pessoa

Como quantificar? Aposta

No exemplo em questão

- Antes do primeiro lançamento: saídas equiprováveis (clássica)
- Após cada lançamento: probabilidades atualizadas
- Após muitos lançamentos: tende ao frequentista

História

- 1740: Thomas Bayes
  "An essay toward solving a problem in the doctrine of chances" na "Philosophical Transactions of the Royal Society of London"
- Solução para o problema inverso (ex.: mesa de bilhar)
- Publicado postumamente por Richard Price
- Ignorando por longo tempo
- Redescoberto e reformulado por Laplace

História

- Condenada pelo mainstream estatístico da época
- Avanços teóricos por De Finet, Jeffrey e Savage
- Adotada durante século XX como alternativa ao frequentismo
- Popularização na década de 90 devido a métodos de Markov Chain Monte Carlo (MCMC)

Aplicações

- Decodificação do Enigma
- Predição de eleições (Nate Silver)
- Filtragem Anti-spam
- Determinação de autoria de documentos
- Investigação forense
- Teologia

Aplicações em Engenharia

- Mais flexível do que o frequentista
- Incorporação de conhecimento prévio
- Sistemas de comunicação
  - Canal
  - Ruído
  - Dados transmitidos
- Tratamento de parâmetros como variáveis aleatórios permite integração de parâmetros auxiliares

3 – O Teorema de Bayes

Motivação: sistema CDMA

- Diversos usuários compartilham um meio
- Canal gera distorções e ruído
- Receptor: analisa sinal recebido e decide a mais provável sequência de dados transmitidos por cada usuário
- Como?

Formalização

- Dados observados: $\mathbf{x}$ (V.A. N-dimensional)
- Distribuição dos dados: $p(\mathbf{x}; \boldsymbol{\theta})$
- $\boldsymbol{\theta}$: parâmetros a serem estimados
- Teorema relaciona distribuição a posteriori com distribuição a priori

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}$$

- Distribuição a priori: $p(\boldsymbol{\theta})$
- Verossimilhança: $p(\mathbf{x}|\boldsymbol{\theta})$
- $p(\mathbf{x}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ (constante de proporcionalidade)

3.1 – Problema 1 revisitado

Uma mulher recebeu o resultado de sua mamografia e avaliação foi anormal. Qual a probabilidade de que ela tenha câncer de mama?

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 %.

Solução

Teorema de Bayes

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

B: mulher tem cancer
A: mamografia anormal

$$p(A) = p(A|B)p(B) + p(A|\overline{B})p(\overline{B})$$

Fazendo as contas:

$$p(A) = 0.8 \times 0.004 + 0.1 \times 0.996 = 0.1028$$

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

4 – Bayesianos vs Frequentistas

Elemento comum

Função de verossimilhança: Probabilidade de os dados $\mathbf{x}$ serem observados quando se assume que os parâmetros corretos são $\boldsymbol{\theta}$

Frequentistas

- Inferência com base apenas na verossimilhança
- $\boldsymbol{\theta}$ é entendido como um parâmetro que aparece na distribuição de $p(\mathbf{x})$ e não como uma variável condicionante

Bayesianos

- Verossimilhança é o elemento de ligação entre a priori e a posteriori (regra de Bayes)
- $\boldsymbol{\theta}$ é entendida como variável que condiciona a distribuição dos dados observados $\mathbf{x}$.

Critério de Máxima Verossimilhança

Ideia: os parâmetros estimados são aqueles que tornam os dados observados os mais prováveis.

Exemplo: De 1000 produtos, 10 são defeituosos. Qual a probabilidade de produto defeituoso?

Critérios bayesianos

- Baseados na posteriori
- O Mínimo Erro Quadrático Bayesiano (BMSE, *Bayeian Minimum Square Error*)
- Máximo *A Posteriori* (MAP, *Maximum A Posteriori*)

4.1 – Examplo: Lançamento de moeda

Deseja-se estimar a probabilidade θ de sair 'cara' no lançamento de uma moeda.

Resposta padrão: $1/2$ (simetria)

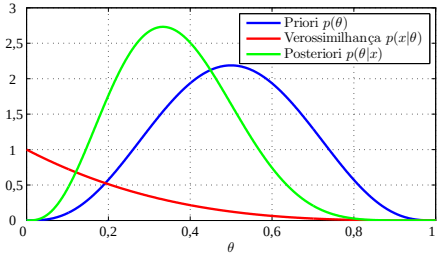
Frequentistas

- 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
- A verossimilhança $p(\mathbf{x}|\theta)$, isto é, a probabilidade de o resultado $\mathbf{x} = [1\ 1\ 1]$ acontecer é obtido considerando cada possível valor de $\theta \in [0\ 1]$
- Supondo independência entre os lançamentos $p(\mathbf{x}|\theta) = (1-\theta)^3$
- Claramente a verossimilhança é maximizada quando $\theta = 0$

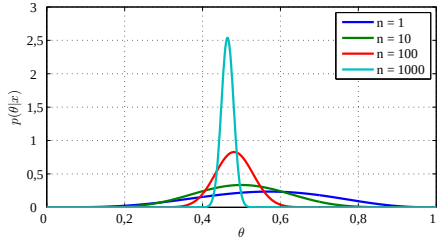
Bayesianos

- Probabilidade *a priori* $p(\theta)$ refletindo sua experiência prévia sobre o lançamento de moedas
- À medida que novos experimentos são feitos, verossimilhança é calculada e combinada com a priri
- Posterioris cada vez mais informativas são geradas

Três lançamentos



Variando número de lançamentos



4.2 – Exemplo 4: Modelo linear generalizado

Útil em diversas aplicações, inclusive comunicações

- Dados observados dependem linearmente de um conjunto de valores desconhecidos a serem estimados
- Ruído aditivo está presente na medida
- Aplicação: sinal recebido em comunicações sem fio com múltiplos percursos e ruído aditivo.

Definição

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

Hipótese

Ruído branco gaussiano de média zero e variância σ_v^2 (possivelmente desconhecida)

Verossimilhança

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

Máxima verossimilhança

Igualando a zero o gradiente do argumento

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

Solução Bayesiana

- Parâmetros $\boldsymbol{\theta}$ tratados como variáveis aleatórias com distribuição *a priori*
- Priori gaussiana com média $\mathbf{m}_{\boldsymbol{\theta}}$ e matriz de covariância $\mathbf{C}_{\boldsymbol{\theta}}$
- Maximização da posteriori gera

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

Análise

- Se os elementos de $\mathbf{C}_\theta$ forem elevados, priori é irrelevante. MAP se aproxima da ML
- Se dimensão dos dados é alta, termos dependendo de $\mathbf{G}$ e $\mathbf{x}$ dominariam a priori. MAP também se aproximaria do ML
- Com poucos dados e conhecimento a priori relevante, soluções diferem

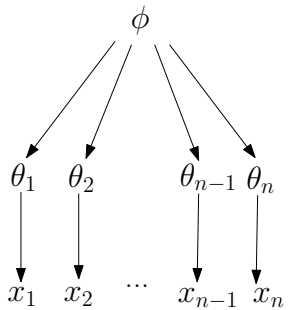
5 – Modelo Bayesiano Hierárquico

Motivação

- Em comunicações sem fio o sinal recebido depende do canal que pode ser caracterizado por um filtro linear com coeficientes desconhecidos
- Bayesiano atribuiria um modelo probabilístico para os coeficientes do canal
- Como os parâmetros desse modelo seriam também desconhecidos, sua descrição depende de outros parâmetros – que passam a ser chamados de hiperparâmetros

Exemplo

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.



Distribuição *a posteriori*

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

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

5.1 – Eliminação de parâmetros

- É comum o vetor de parâmetros conter elementos que não são de interesse para estimação.
- Escola Bayesiana permite integrar parâmetros *nuisance*
- Particionando $\boldsymbol{\theta}$ 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 fica

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

5.2 – Distribuição a priori

- Quantifica conhecimento prévio
- Especificação é difícil na prática
- *Trade-off* entre realismo e tratabilidade matemática

Prioris conjugadas

- Prioris com a mesma estrutura algébrica da verossimilhança
- Produto entre ambas tem forma de simples manipulação
- Exemplo: se a verossimilhança é gaussiana, a escolha de uma priori também gaussiana gera uma posteriori gaussiana

Revisitando modelo linear generalizado

Assumindo priori para $\boldsymbol{\theta}$ gaussiana com média $\mathbf{m}_\theta$ e matriz de covariância $\mathbf{C}_\theta$, a posteriori será:

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

Variância do ruído σ_v^2 pode ser descrita probabilisticamente. Como σ_v^2 aparece de forma similar a uma distribuição gama inversa, sua priori conjugada é uma gama-inversa $p(\sigma_v^2|\alpha_v, \beta_v)$.

A posteriori fica:

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

Priori não-informativa

- Priori não-informativa idealmente não contém qualquer informação sobre os possíveis valores dos parâmetros
- Escolha aparentemente mais natural: uniforme
- Problemática em espaços contínuos, uma vez que a distribuição uniforme não é invariante a transformações bijetivas de variáveis
- Se $\boldsymbol{\theta}$ é uniforme, a distribuição de $\boldsymbol{\phi} = f(\boldsymbol{\theta})$ não será mais uniforme
- Exemplo: Volume de uma esfera

A priori de Jeffreys

Invariante à transformações uma-para-um

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

onde $\mathbf{I}(\boldsymbol{\theta})$ é a matriz informação de Fisher

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

No caso de modelos invariantes com relação à escala

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

6 – Simulação Estocástica através MCMC

- Se distribuições são multivariáveis e multi-modais, técnicas clássicas de otimização são inadequadas
- Modelagem realística requer técnicas numéricas mais sofisticadas
- Exemplos: Algoritmo EM (*Expectation-Maximization*), Amostrador de Gibbs, Algoritmos de Metropolis-Hastings (MH)

Técnicas de Monte Carlo

- Geração de amostras i.i.d.de uma certa distribuição
- Aproximação de alguma característica da distribuição
- Partindo de amostras $X = \{x^{(1)}, \ldots, x^{(N)}\}$ de $p(x)$, aproxima-se

$$I(f) = \int f(x)p(x)dx$$

por

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

- Exemplo: média amostral

Exemplo

- Cálculo da área de um círculo unitário
- Seja $p(x)$ uma distribuição uniforme no quadrado de vértices $(1, 1)$, $(1, -1)$, $(-1, 1)$ e $(-1, -1)$
- $f(x)$ uma função indicadora que retorna 1 se x estiver no interior do círculo e 0 caso contrário
- Gerando $N = 1000$ amostras de $p(x)$, encontramos $I_N(f) \approx 3.13988$

Cadeias de Markov

- Amostragem de $p(x)$ nem sempre é trivial
- Cadeias de Markov especialmente projetadas permitem amostragem indireta
- Definição: processo aleatório discreto no tempo que apresenta em que o estado atual da cadeia depende unicamente do estado imediatamente anterior
- 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 $X^{(n)}$

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

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

Espaço de estados discreto

- S é um conjunto contável
- $S = \{s_1, \ldots, s_N\}$
- Matriz de transição $\mathbf{T}_n$

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

- Soma das colunas é unitária (matriz estocástica)
- Ao menos um autovalor unitário
- Se $T_n(i|j) > 0$: todos os demais autovetores distintos e menores que 1

Distribuição de probabilidade no instante n, $P_n(i)$

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

Vetorialmente

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

em que $\mathbf{P}_n = [P_n(1) \dots P_n(N)]^T$.
Se $\mathbf{T}_n$ é independente de n,

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

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

- 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
- Aperiodicidade: A cadeia não ficar presa em ciclos

Distribuição invariante

Def.: Uma distribuição é dita invariante se permanece fixa sob a aplicação da matriz de transição
Se $\pi(.)$ é invariante:

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

Podemos escrever $\boldsymbol{\pi} = \mathbf{T}\boldsymbol{\pi}$, donde vemos que $\boldsymbol{\pi}$ é um autovetor associado ao autovalor $\lambda = 1$.

Detailed Balance

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:

$$\pi(j)T(i|j) = \pi(i)T(j|i).$$

Nessa situação, pode-se mostrar que $\pi(i)$ é distribuição invariante

Ergodicidade

- Além de garantir que $\pi(i)$ é 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 e a cadeia é ergódica.
- Para ergodicidade, a cadeia precisa ser aperiódica e irredutível.
- No caso discreto, basta termos os autovalores de $\mathbf{T}$ todos distintos,
- Escrevemos a distribuição inicial usando os autovetores de $\mathbf{T}$ como base
- Calculamos a expressão $\mathbf{T}^n\mathbf{P}_0$

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

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

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

Caso contínuo

A propriedade de Markov torna-se

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

Núcleo de transição $K_n(x|y)$ no instante n

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

$p_{X^{(n)}}(x)$: densidade de probabilidade de $X^{(n)}$.
Distribuição do estado da cadeia no instante n

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

Detailed Balance

No caso de a cadeia ser homogênea, K_n independente de n

$$\int_A \int_B K(x|y)\pi(y)dydx = \int_B \int_A K(y|x)\pi(x)dxdy,$$

para quaisquer conjuntos A e B pertencentes a S

- Como no caso discreto, *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

Amostrador de Gibbs

- Proposta: Geman, S.; Geman, D. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". IEEE Transactions on Pattern Analysis and Machine Intelligence
- 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.

Mais formalmente

- Seja $\pi(\boldsymbol{\theta})$ a distribuição conjunta da qual se deseja obter amostras
- A variável $\boldsymbol{\theta}$ é 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:

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

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

$$\vdots$$

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

Análise

- 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
- Supondo que a distribuição da cadeia de Markov no final da iteração $(i-1)$ é $\pi(\boldsymbol{\theta})$, isto é: $p(\boldsymbol{\theta}^{(i-1)}) = \pi(\boldsymbol{\theta}^{(i-1)})$

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

Algoritmo de Metropolis-Hastings

- Proposto por Metropolis et al (1953) e generalizado por Hastings (1970)
- Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. (1953). "Equations of State Calculations by Fast Computing Machines". Journal of Chemical Physics
- Hastings, W.K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". Biometrika
- Útil quando distribuições condicionais são difíceis de amostrar
- Ideia: 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

Definição formal

- Uma amostra x^* é obtida a partir de uma distribuição proposta, $q(x^*|x^{(i)})$, em que $x^{(i)}$ é o estado atual da cadeia de Markov
- Essa amostra é aceita com uma probabilidade α

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

Análise

- Se a amostra gerada for aceita, o novo estado da cadeia é $x^{i+1} = x^*$
- Caso contrário, a cadeia permanece no seu estado atual, isto é, $x^{(i+1)} = x^{(i)}$
- Núcleo de transição

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

em que

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

- Satisfaz a condição *detailed balance*: $\pi(x)$ é uma distribuição invariante
- Como o algoritmo sempre permite a rejeição: cadeia é aperiódica.
- Para irredutibilidade, o suporte de $q(\cdot)$ deve incluir suporte de $\pi(\cdot)$

Discussão

- 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, $q(x^*|x^{(i)}) = N(x^*|x^{(i)}, \sigma_q^2 I)$
- Se $q(\cdot)$ é muito estreita, apenas estados próximos ao máximo de $\pi(x)$ são visitados
- Se $q(\cdot)$ é muito ampla, o percentual de amostras rejeitadas é muito alto e, consequentemente, as amostras geradas serão altamente correlacionadas entre si, e a convergência será lenta

7 – Deteção de sinais em canais seletivos em frequência usando inferência Bayesiana

- Esquema de modulação digital BPSK (*Binary Phase Shift-Keying*)
- Símbolos binários x_1, x_2, ..., x_N, com $x_k \in \{+1, -1\}$
- Canal seletivo em frequência: sinal recebido y_n é distorcido e pode ser escrito como combinação linear dos dados enviados deslocadas no tempo, adicionada ao ruído do canal

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

- $\mathbf{h} = \{h_1, \ldots, h_L\}$ representa a resposta ao impulso do canal de comprimento L
- $\mathbf{v} = \{v_1, \ldots, v_N\}$ são amostras i.i.d de ruído branco gaussiano de média zero e variância σ_v^2

Abordagem Bayesiana

- Problema consiste em estimar a sequência enviada $\mathbf{x} = \{x_1, \ldots, x_N\}$ com base no sinal recebido $\mathbf{y} = \{y_1, \ldots, y_N\}$
- Tratamos como *nuisance* o conjunto de parâmetros $\boldsymbol{\theta} = \{h_0, \ldots, h_L, \sigma_v^2\}$
- Necessária obtenção da distribuição *a posteriori* das quantias desconhecidas – dados $\mathbf{x}$ e os parâmetros *nuisance* $\boldsymbol{\theta}$.

Teorema de Bayes

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

onde

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

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

Verossimilhança

- Situação parecida com o modelo linear geral
- Verossimilhança dada por:

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

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

Posteriori

- Substituindo as prioris e a verossimilhança no teorema de Bayes a posteriori é obtida.
- Como queremos estimar apenas os dados enviados, precisamos calcular a distribuição condicional dos dados enviados sendo conhecidos os dados recebidos:

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

Solução numérica

- A integral acima não pode ser resolvida analiticamente
- 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
- 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.

Amostrador de Gibbs

Descrição de alto nível

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**

Amostragem de $\mathbf{x}$

- Amostragem do vetor $\mathbf{x}$ inteiro 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 $\mathbf{h}$

- Para calcularmos a distribuição condicional de $\mathbf{h}$ é conveniente escrever a equação do canal de uma forma alternativa explicitando a dependência dos dados com o parâmetro $\mathbf{h}$

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

em que $\mathbf{X}$ é formada por elementos de $\mathbf{x}$

- A condicional total de $\mathbf{h}$ é dada por:

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

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

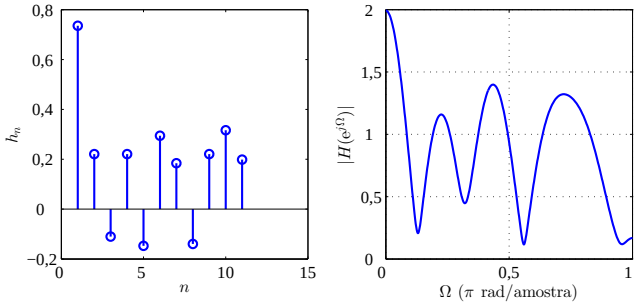
Amostragem de σ_v^2

- Priori do tipo gama inversa para a variância σ_v^2 gera uma posteriori do mesmo formato com parâmetros modificados
- Posteriori

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

Simulações

- Canal severamente seletivo em frequência
- Escolhemos o canal de comprimento $L = 11$ com resposta ao impulso

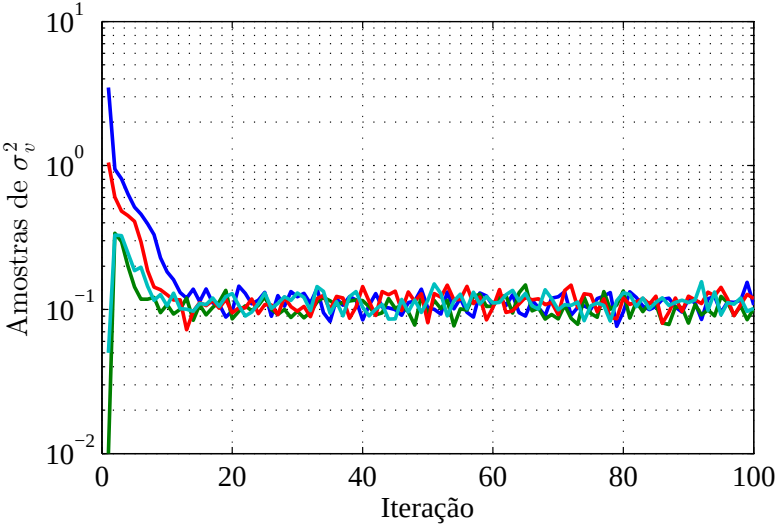


- Variância do ruído $\sigma_v^2 = 0.1$ correspondendo a uma SNR de 10 dB

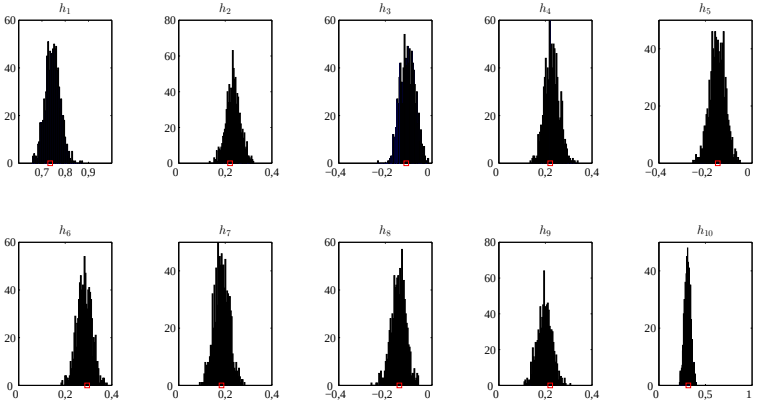
Escolhas

- Inicializamos os parâmetros com $\mathbf{x}^0 = \text{sign}(\mathbf{y})$, $\sigma_v^{2(0)} \sim \text{IG}(\sigma_v^2|2,1)$
- O algoritmo é executado por 1000 iterações partindo de vários valores iniciais para σ_v^2.

Análise de convergência



Histogramas após convergência



Discussão

- Convergência ocorre após algumas dezenas de iterações
- Algoritmo estima satisfatoriamente os parâmetros do modelo
- A partir dos valores amostrados para $\mathbf{x}$ podemos facilmente calcular a estimativa de máximo *a posteriori* para cada símbolo enviado
- 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 %.

Variações do problema

Tipo de sistema

- QAM: Punskaya et al, "Particle Filters for Demodulation of M-Ary Modulated Signals in Noisy Fading Communication Channels", ICASSP, 2000
- CDMA: Farhang-Boroujeny, "Markov Chain Monte Carlo Algorithms for CDMA and MIMO Communication Systems", IEEE T. on Sig. Proc., 2006.
- OFDM: Prasad et al, "Joint Channel Estimation and Data Detection in MIMO-OFDM Systems: A Sparse Bayesian Learning Approach", IEEE Trans. on Sig. Proc., 2015
- MIMO: Wen et al, "Channel Estimation for Massive MIMO using Gaussian-Mixture Bayesian Learning", IEEE T. on Wireless Comm, Março 2015. Ling et al, "On Bayesian Channel Estimation and FFT-Based Symbol Detection in MIMO Underwater Acoustic Communications", IEEE T. on Oceanic Engineering, 2014.

Variações do problema

Tipo de canal

- Efeito Doppler: canal variante no tempo
  - Processamento frame a frame
  - Monte Carlo Sequencial (Particle filtering)
  - Nevat et al, "Joint Channel and Doppler Offset Estimation in Dynamic Cooperative Relay Networks", IEEE T. on Wireless Comm.", 2014
- Canal não-linear
  - Distribuições ficam complicadas (não-Gaussianas)
  - Aproximação gaussiana com Metropolis-Hastings
  - Particle filtering
  - Li et al, "A Bayesian Approach for Nonlinear Equalization and Signal Detection inMillimeter-Wave Communications", IEEE T. on Wireless Comm., Julho, 2015.

Variações do problema

Ruído

- Impulsivo
- Não-Gaussiano
- Variante no tempo

Para aprender mais

Livros e tutoriais

- Yang et al, "Fifty Years of MIMO Detection: The Road to Large-Scale MIMOs", IEEE Comm. Surveys & Tutorials, 2015
- X. Wang e V. Poor, "Wireless Communication Systems: Advanced Techniques for Signal Reception", 2002
- "Advances in Multiuser Detection", Editada por M. Honig, 2009
- Bajwa et al, "Compressed Channel Sensing: A New Approach to Estimating Sparse Multipath Channels", Proceedings of the IEEE, 2010.

Conclusões

- Neste mini-curso vimos os fundamentos de inferência Bayesiana e das técnicas de simulação estatística
- Enfoque no problema de deteção de sinais em sistemas de comunicações
- Estudo de caso: canal seletivo em frequência, invariante no tempo com ruído gaussiano
- Embora simplificado, esse modelo nos permite entender os principais aspectos da análise Bayesiana
  - Escolha das distribuições *a priori* para os parâmetros
  - Cálculo da verossimilhança
  - Cálculo da distribuição *a posteriori* pelo teorema de Bayes
  - Solução numérica do problema de estimação através de MCMC

Conclusões

- 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
- Inicialização favorável pode acelerar convergência

Conclusões

- 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