Análise de Dados em Cosmologia: Métodos Bayesianos

Laerte Sodré Jr.

September 22, 2011

A teoria das probabilidades é o senso comum reduzido ao cálculo

Pierre-Simon Laplace

Tópicos:

1. Introdução
2. Probabilidades
3. As regras da teoria das probabilidades
4. Métodos bayesianos e análise de dados
5. Inferência dos parâmetros de um modelo
6. Inferência dos parâmetros de uma distribuição gaussiana
7. Inferência dos parâmetros de uma distribuição gaussiana bidimensional
8. Inferência dos parâmetros de uma distribuição gaussiana multidimensional
9. Aplicação : a estatística de Riess et al. (1998)
10. Comparação de modelos

1. Introdução : O que são probabilidades?

- duas grandes correntes sobre a *natureza* das probabilidades:
  - bayesiana: medida da plausibilidade de uma proposição
  - frequentista: frequências de ocorrência de eventos (em vários experimentos ou *ensemble* de sistemas estatisticamente equivalentes)
- território polêmico: vejam o que diz o *Numerical Recipes* e procurem no Google por *bayesian versus frequentist*

1. Introdução : O que são probabilidades?

- probabilidade frequentista:
  o número de vezes que um evento ocorre dividido pelo número total de tentativas, no limite de um número infinito de tentativas
- isso não é bom:
  - não comporta eventos únicos ou não repetíveis
  - lida com as propriedades assintóticas (“no limite...”)
- o enfoque bayesiano não depende disso!

1. Introdução : Cox e as regras para um raciocínio consistente

- Cox (1946): estuda as regras para um *raciocínio consistente*
- Como expressar nossas crenças relativas na verdade de várias proposições como
  - (a) amanhã o Sol vai nascer
  - (b) vai chover amanhã
  - (c) esta moeda é viciada
- ele demonstra que um requisito mínimo é a *transitividade*:
  se acredito em (a) mais que em (b) e em (b) mais do que em (c), então necessariamente devo acreditar mais em (a) que em (c)
- a transitividade pode ser implementada atribuindo-se um número real a cada proposição, tal que, quanto maior o número, mais acreditamos nessa proposição
- Cox mostra que o raciocínio consistente só é possível se os números reais associados às proposições obedecerem às regras usuais da probabilidade

- probabilidade: medida da plausibilidade de uma proposição quando não se pode estabelecer com certeza se ela é verdadeira ou falsa
  - $P(x)$: número entre 0 e 1 que mede o grau com que a proposição x é verdadeira (com 0 falsa e 1 verdadeira)
- quando se tem um contínuo de proposições, $P(x)$ fica uma *função de densidade de probabilidades* (pdf):
  $P(x)dx$: número entre 0 e 1 que mede o grau de plausibilidade de que uma certa variável x esteja entre x e $x + dx$

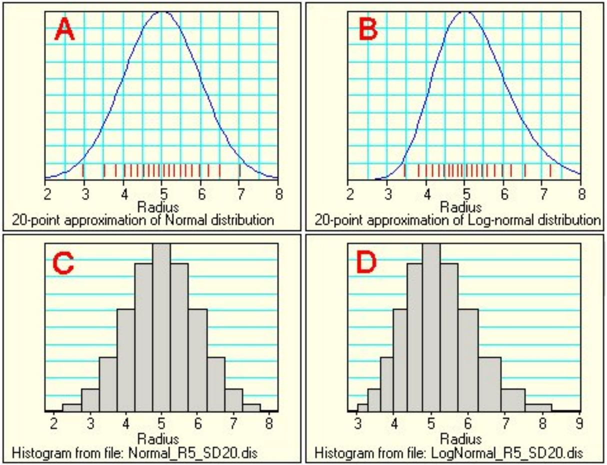


Figure: exemplos de distribuições de probabilidade contínuas (acima) e discretas (abaixo)

2. Probabilidades e valores esperados

- valor esperado de uma função $Q(x)$ com respeito a uma pdf $P(x)$:

$$< Q >= \int Q(x)P(x)dx$$

- para um conjunto finito de N pontos $\{x_i\}$, tirados da distribuição $P(x)$, o valor esperado pode ser aproximado pela média dos dados:

$$< Q >= \int Q(x)P(x)dx \simeq \frac{1}{N}\sum_{i=1}^{N} Q(x_i)$$

- essa última expressão é a base de métodos tipo Monte Carlo

2. Probabilidades condicionais

- probabilidades são, normalmente, *condicionais*, isto é, dependem ou podem depender de outras proposições
  - $P(x|y)$: lê-se probabilidade de x *dado* y
  - $P(x|y)$ mede a plausibilidade da proposição x se a proposição y é admitida como verdadeira
  - toda probabilidade é condicional a alguma coisa: não existe probabilidade absoluta
    a probabilidade de chover amanhã vai depender do clima de hoje
- tudo o que estiver a direita da barra "|" é suposto conhecido

a distribuição gaussiana:

- a distribuição gaussiana (ou normal) é frequentemente usada como modelo para a distribuição de erros de uma medida:

$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \equiv \mathrm{Normal}(x;\mu,\sigma)$$

  onde μ é o valor médio e σ uma medida do erro nas medidas
- a probabilidade de x estar dentro de $\pm 1\sigma$ de x_0 é 0.683, dentro de $\pm 2\sigma$ é 0.954 e dentro de $\pm 3\sigma$ é 0.997

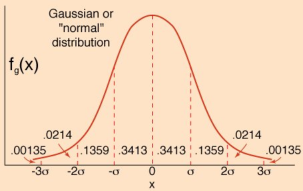


Figure: distribuição gaussiana; mostra-se as frações da área em intervalos de σ

2. Probabilidade conjunta

- $P(x, y|H)$: *probabilidade conjunta* de duas proposições x e y, dado que uma certa proposição H é verdadeira
  - H pode, por exemplo, representar as hipóteses que sustentam estas proposições
  - algumas vezes, para simplificar a notação, se abole H

2. As regras da teoria das probabilidades

- as 2 regras fundamentais são:
  - regra da soma:
    $$P(x|H) + P(\bar{x}|H) = 1$$
    onde $\bar{x}$ representa a probabilidade de x ser falso
  - regra do produto:
    $$P(x, y|H) = P(x|y, H) \times P(y|H)$$

2. A regra da soma

- aplicações:
  - regra da soma:
    $$P(x|H) + P(\bar{x}|H) = 1$$
  - generalização para um conjunto de proposições x_k mutuamente exclusivas e exaustivas:
    $$\sum_k P(x_k|H) = 1$$
    que é a regra de normalização das probabilidades das proposições x_k
  - no caso contínuo:
    $$\int P(x|H)dx = 1$$

2. Proposições independentes

- aplicações:
  - vamos supor que x e y são independentes
  - então,

$$P(x|y,H) = P(x|H)$$
$$P(y|x,H) = P(y|H)$$

e, portanto,

$$P(x,y|H) = P(x|y,H)P(y|H) = P(x|H)P(y|H)$$

isto é, a probabilidade conjunta de duas proposições independentes é o produto das probabilidades de cada uma

- corolário 1: o *teorema de Bayes*
  - da regra do produto vem que

$$P(x, y|H) = P(x|y, H) \times P(y|H)$$

$$P(y, x|H) = P(y|x, H) \times P(x|H)$$

  - como $P(x, y|H) = P(y, x|H)$, temos

$$P(x|y, H) = \frac{P(y|x, H) \times P(x|H)}{P(y|H)},$$

    que é o teorema de Bayes

2. O teorema de Bayes

O reverendo Thomas Bayes

O teorema foi descoberto por volta de 1740

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

2. A marginalização

- corolário 2: a *marginalização*
  - quando se tem um conjunto y_k de proposições mutuamente exclusivas e exaustivas,

$$P(x|H) = \sum_k P(x, y_k|H)$$

  - ou, no caso contínuo,

$$P(x|H) = \int P(x, y|H)dy$$

    onde $P(x, y|H)$ é uma função de densidade de probabilidades (pdf)

- a marginalização é um jeito excelente de "sumir" com parâmetros indesejados, conhecidos como *nuisance parameters*

2. A marginalização - demonstração

- corolário 2: a *marginalização* - demonstração
  - da regra do produto,

$$P(x, y|H) = P(y, x|H) = P(y|x, H) \times P(x|H)$$

$$P(x, \bar{y}|H) = P(\bar{y}, x|H) = P(\bar{y}|x, H) \times P(x|H)$$

  - logo,

$$P(x, y|H) + P(x, \bar{y}|H) = [P(y|x, H) + P(\bar{y}|x, H)] \times P(x|H) = P(x|H)$$

ou

$$P(x|H) = P(x, y|H) + P(x, \bar{y}|H)$$

  - no caso de um conjunto y_k de proposições mutuamente exclusivas e exaustivas,

$$P(x|H) = \sum_k P(x, y_k|H)$$

2. A marginalização - demonstração

- corolário 2: a *marginalização* - demonstração
  - conjunto y_k de proposições mutuamente exclusivas e exaustivas:
  $$P(x|H) = \sum_k P(x, y_k|H)$$
  - no caso contínuo,
  $$P(x|H) = \int P(x, y|H) dy$$
  e vale a normalização (definição de p.m.e.e.):
  $$\int P(y|x, H) dy = 1$$

3. Métodos bayesianos e análise de dados

- consideremos um conjunto de dados D que queremos explicar com um modelo H que tem alguns parâmetros x
- o teorema de Bayes pode ser escrito como

$$P(x|D,H) = \frac{P(D|x,H) \times P(x|H)}{P(D|H)}$$

- cada termo tem um nome:
  - $P(x|D,H))$: **posterior**, ou probabilidade *a posteriori* dos parâmetros
  - $P(D|x,H)$: função de **verossimilhança** dos dados (likelihood)
  - $P(x|H)$: **prior**, ou probabilidade *a priori* dos valores dos parâmetros
  - $P(D|H)$: **evidência** dos dados

$$P(D|H) = \int P(D|x,H) \times P(x|H) dx$$

posterior likelihood prior

$$P(H|d,I) = \frac{P(d|H,I)P(H|I)}{P(d|I)}$$

evidence

- **Prior**: what we know about H (given information I) before seeing the data
- **Likelihood:** the probability of obtaining data d if hypothesis H is true
- **Posterior:** our state of knowledge about H after we have seen data d
- **Evidence**: normalization constant (independent of H), important for model comparison (see later)

3. Métodos bayesianos e análise de dados

- Os métodos bayesianos permitem uma abordagem lógica de dois problemas importantes da ciência empírica:
  - estimativa de parâmetros de modelos
  - comparação de modelos

3. Métodos bayesianos e análise de dados: estimativa de parâmetros de modelos

- estimativa de parâmetros de modelos:

$$P(x|D,H) \propto P(D|x,H) \times P(x|H)$$

  - os parâmetros são selecionados maximizando o posterior
  - a evidência não é importante nesse tipo de problema
  - para um prior uniforme ($P(x|H) = \text{const.}$) a melhor estimativa é a que maximiza a verossimilhança
  - erros: podem ser estimados como níveis de confiança do posterior

3. Métodos bayesianos e análise de dados: comparação de modelos

- comparação de modelos
  - considere dois modelos, H_1 e H_2, para explicar um conjunto de dados D; que modelo deve ser preferido?
  - teorema de Bayes:

$$P(H_i|D, I) = \frac{P(D|H_i, I) \times P(H_i|I)}{P(D|I)}$$

    onde I representa a informação ainda mais fundamental que temos sobre o problema
  - logo

$$\frac{P(H_1|D, I)}{P(H_2|D, I)} = \frac{P(D|H_1, I)}{P(D|H_2, I)} \times \frac{P(H_1|I)}{P(H_2|I)}$$

3. Métodos bayesianos e análise de dados: comparação de modelos

- comparação de modelos
  - considere dois modelos, H_1 e H_2, para explicar um conjunto de dados D; que modelo deve ser preferido?
  - temos que
$$\frac{P(H_1|D,I)}{P(H_2|D,I)} = \frac{P(D|H_1,I)}{P(D|H_2,I)} \times \frac{P(H_1|I)}{P(H_2|I)}$$
  - se supusermos que todos os modelos são igualmente prováveis *a priori*,
$$P(H_1|I) = P(H_2|I)$$
  - então,
$$\frac{P(H_1|D,I)}{P(H_2|D,I)} = \frac{P(D|H_1,I)}{P(D|H_2,I)}$$
logo, a razão entre a probabilidade dos modelos é diretamente proporcional à evidência de cada um
  - (se os modelos não são igualmente prováveis, é necessário considerar o prior de cada um)

3. Métodos bayesianos e análise de dados: comparação de modelos

- comparação de modelos: nomenclatura
  - odds:
$$\frac{P(H_1|D,I)}{P(H_2|D,I)}$$
  - fator de Bayes:
$$\frac{P(D|H_1,I)}{P(D|H_2,I)}$$
- comparação de modelos: *odds* $=$ *Bayes factor*

$$\frac{P(H_1|D,I)}{P(H_2|D,I)} = \frac{P(D|H_1,I)}{P(D|H_2,I)}$$

3. Métodos bayesianos e análise de dados: comparação de modelos

| lnB |	relative odds	favoured model's probability	Interpretation
< 1.0	$< 3{:}1$	< 0.750	not worth mentioning
< 2.5	$< 12{:}1$	0.923	weak
< 5.0	$< 150{:}1$	0.993	moderate
> 5.0	$> 150{:}1$	> 0.993	strong

Figure: A escala de Jeffreys (modificada por Trotta) para avaliar a evidência.

4. Inferência dos parâmetros de um modelo

- estimativa de parâmetros de modelos:

$$P(x|D,H) \propto P(D|x,H) \times P(x|H)$$

  a evidência não é importante neste tipo de problema pois não depende de x

- vamos resumir o resultado da determinação de um parâmetro com dois números:
  a melhor estimativa e uma medida de sua confiabilidade
  (o "erro")

4. Inferência dos parâmetros de um modelo

- estimativa de parâmetros de modelos:

$$P(x|D,H) \propto P(D|x,H) \times P(x|H)$$

- como o posterior é uma medida do nosso grau de plausibilidade no valor de um parâmetro, a melhor estimativa é o valor x_0 que maximiza o posterior:

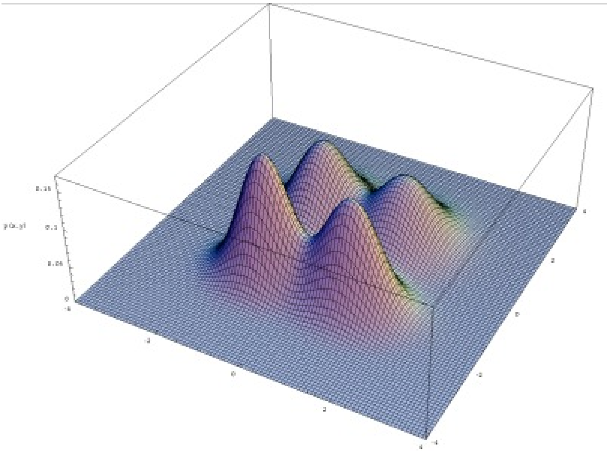
$$\left(\frac{dP(x|D,H)}{dx}\right)_{x_0} = 0$$

e a condição de máximo

$$\left(\frac{d^2P(x|D,H)}{dx^2}\right)_{x0} < 0$$

4. Inferência dos parâmetros de um modelo

- cuidado: o posterior pode ser multimodal!



5. Inferência dos parâmetros de uma distribuição gaussiana

problema:

- temos um conjunto de dados sobre uma variável x, $D = \{x_k\}$ com $k = 1, ..., N$; vamos supor que estão distribuídos como uma gaussiana de largura σ; qual é o melhor valor de μ e qual é nossa confiança nele?
- modelo para os dados:
  $x_k = \mu +$ ruído gaussiano de largura σ

$$P(x_k|\mu, \sigma, H) \propto \exp\,[-\frac{(x_k - \mu)^2}{2\sigma^2}]$$

5. Inferência dos parâmetros de uma distribuição gaussiana

- teorema de Bayes:

$$P(\mu|D,\sigma,H) = \frac{P(D|\mu,\sigma,H)P(\mu|\sigma,H)}{P(D|H)}$$

- a verossimilhança dos dados é

$$P(D|\mu,\sigma,H) = P(\{x_k\}|\mu,\sigma,H)$$

- vamos supor que as medidas $\{x_k\}$ são independentes; nesse caso, a probabilidade dos dados é o produto das probabilidades das medidas:

$$P(\{x_k\}|\mu,\sigma,H) = \prod_{k=1}^{N} P(x_k|\mu,\sigma,H)$$

o prior

- vamos atribuir uma pdf uniforme para o prior de μ:

$$P(\mu|\sigma, H) = P(\mu|H) = (\mu_{max} - \mu_{min})^{-1}$$

se $\mu_{min} < \mu < \mu_{max}$, e

$$P(\mu|\sigma, H) = P(\mu|H) = 0$$

fora desse intervalo

- usando o teorema de Bayes e considerando o logaritmo do posterior,

$$L = \ln P(\mu|D,\sigma,H) = \text{constante} - \sum_{k=1}^{N} \frac{(x_k-\mu)^2}{2\sigma^2}$$

  onde a constante inclui todos os termos que não dependem de μ
- o posterior é nulo fora do intervalo $[\mu_{min}, \mu_{max}]$: pontos fora do intervalo não são considerados- vamos supor que nenhum ponto é excluído

5. Inferência dos parâmetros de uma distribuição gaussiana

- logaritmo do posterior:

$$L = \ln P(\mu|D, \sigma, H) = \mathrm{constante} - \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2}$$

- a melhor estimativa, μ_0, é a que maximiza o posterior e, portanto,

$$\left.\frac{dL}{d\mu}\right|_{\mu_0} = \sum_{k=1}^{N} \frac{(x_k - \mu_0)}{\sigma^2} = 0$$

- ou

$$\mu_0 = \frac{1}{N} \sum_{k=1}^{N} x_k$$

  isto é, a melhor estimativa é igual à média aritmética dos valores das medidas

- a melhor estimativa não depende dos erros (σ)

5. Inferência dos parâmetros de uma distribuição gaussiana

- nossa confiança em μ_0 depende da largura do posterior, cuja variancia é

$$< (\mu - \mu_0)^2 >= \int (\mu - \mu_0)^2 P(\mu|D, H) d\mu = \frac{\sigma^2}{N}$$

- a raiz quadrada da variancia é o *desvio padrão* ou a raiz quadrada do *desvio quadrático médio* (rms) de μ
- vimos que a largura do posterior está associada ao inverso da derivada segunda de L:

$$\left.\frac{d^2 L}{d\mu^2}\right|_{\mu_0} = -\sum_{k=1}^{N} \frac{1}{\sigma^2} = -\frac{N}{\sigma^2}$$

- temos que

$$< (\mu - \mu_0)^2 >= \frac{\sigma^2}{N}$$

- vamos representar o "erro" na inferência pelo desvio padrão (raiz quadrada da variancia)
- o resultado da inferência pode então ser expresso como

$$\mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}}$$

isto é, a incerteza na estimativa diminui com a raiz quadrada do número de dados

5. Inferência dos parâmetros de uma distribuição gaussiana

- papel do prior
  - se o intervalo $[\mu_{min}, \mu_{max}]$ cobre todos os dados, a inferência não depende de μ_{min} e μ_{max} e maximizar o posterior é equivalente a maximizar a verossimilhança
  - método da máxima verossimilhança: equivalente ao uso de um prior uniforme, não informativo
  - nesse caso o valor de μ_0 depende só dos dados
  - mas imagine agora que você ache que μ_0 está num intervalo relativamente estreito: nesse caso sua inferência vai depender do prior

5. Inferência dos parâmetros de uma distribuição gaussiana

- o enfoque bayesiano é muitas vezes criticado por usar priores
- algumas críticas:
  - os priores são subjetivos
  - a teoria não especifica os priores
  - diferentes priores levam a diferentes resultados
- respostas:
  - os priores codificam o estado de conhecimento do cientista sobre a hipótese
  - explicitando os priores força a explicitar as hipóteses
  - se os resultados dependem muito dos priores é porque os dados não são suficientemente "explicativos"

6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

- vamos supor que nosso modelo tem 2 parâmetros, X e Y (pode ter mais, mas vamos supor que os outros estão marginalizados)
- o melhor valor dos parâmetros (X_0, Y_0) é obtido maximizando o posterior ou seu logaritmo:

$$\left.\frac{\partial L}{\partial X}\right|_0 = 0 \quad e \quad \left.\frac{\partial L}{\partial Y}\right|_0 = 0$$

- vamos aproximar L por uma gaussiana em torno de (X_0, Y_0):

$$L = L(X_0, Y_0) + \frac{1}{2}\left[\left(\frac{\partial^2 L}{\partial X^2}\right)_0 (X - X_0)^2 + \left(\frac{\partial^2 L}{\partial Y^2}\right)_0 (Y - Y_0)^2 \right.$$

$$\left. +2\left(\frac{\partial^2 L}{\partial X \partial Y}\right)_0 (X - X_0)(Y - Y_0)\right] + \ldots =$$

$$= L_0 + Q + \ldots$$

6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

- aproximação gaussiana em torno de (X_0, Y_0):

$$L = L_0 + Q$$

onde Q, em notação matricial, fica:

$$Q = -\mathbf{W}\ \mathbf{F}(\mathbf{W_0})\ \mathbf{W}^T$$

- $\mathbf{W}$ é um vetor linha com os parâmetros, $\mathbf{F}(\mathbf{W_0})$ a matriz 2×2 das derivadas segundas e o superscrito ^{T} refere-se à matriz transposta
- $\mathbf{F}$ é às vezes chamada de Matriz de Fisher ou Hessiano

6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

- aproximação gaussiana de L em torno de (X_0, Y_0):

$$L = L_0 + Q$$

onde

$$Q = \mathbf{W}\ \mathbf{F}(\mathbf{W_0})\ \mathbf{W}^T$$

$$\mathbf{W} = [X - X_0 \quad Y - Y_0]$$

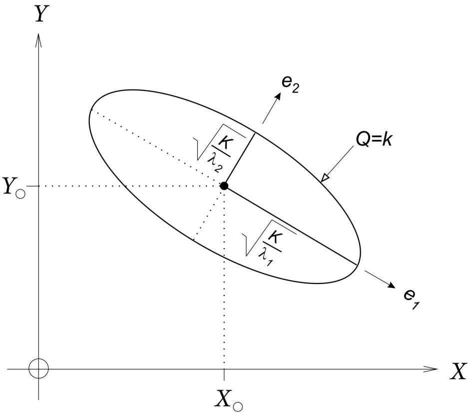
$$\mathbf{F}(\mathbf{W_0}) = -\begin{bmatrix} A & C \\ C & B \end{bmatrix}$$

e

$$A = \left(\frac{\partial^2 L}{\partial X^2}\right)_0 \quad B = \left(\frac{\partial^2 L}{\partial Y^2}\right)_0 \quad C = \left(\frac{\partial^2 L}{\partial X \partial Y}\right)_0$$

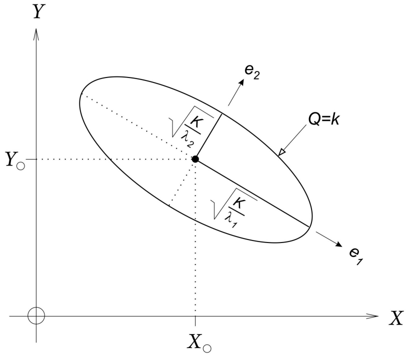
6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

- contorno de $Q = k =$ constante no plano (X, Y): elipses
- ao longo de uma elipse $Q = k$ o posterior é constante



6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

- centro da elipse: (X_0, Y_0)
- orientação , excentricidade e tamanho dependem da matriz $\mathbf{F}(\mathbf{W_0})$, em particular de seus autovetores $\mathbf{e_1}$ e $\mathbf{e_2}$ e autovalores λ_1 e λ_2 autovalores λ



6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

- os autovetores e_1 e e_2 têm componentes (x, y) que são soluções da equação

$$\begin{bmatrix} A & C \\ C & B \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix}$$

  com os respectivos autovalores λ_1 e λ_2

- se (X_0, Y_0) é um máximo (e não um mínimo ou ponto de sela), então λ_1 e λ_2 devem ser negativos, ou

$$A < 0 \quad B < 0 \quad AB > C^2$$

  se $C = 0$ a elipse está alinhada com um dos eixos

6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

- se estamos interessados apenas em X, podemos marginalizar Y:
$$P(X|D,H) = \int P(X,Y|D,H)dY$$
com $P(X,Y|D,H) \propto exp(L) \propto \exp(-Q/2)$
- pode-se verificar que nesse caso temos que
$$P(X|D,H) \propto \exp\left[\frac{1}{2}\left(\frac{AB-C^2}{B}\right)(X-X_0)^2\right]$$
- a melhor estimativa de X é X_0 e o erro associado é
$$\sigma_x = \sqrt{\frac{-B}{AB-C^2}}$$

6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

- erro em X_0:

$$\sigma_x = \sqrt{\frac{-B}{AB - C^2}}$$

- erro em Y_0:

$$\sigma_y = \sqrt{\frac{-A}{AB - C^2}}$$

- por outro lado, a *covariancia* entre X e Y é

$$< (X-X_0)(Y-Y_0) >= \int (X-X_0)(Y-Y_0)P(X,Y|D,H)dXdY =$$

$$= \sigma_{xy}^2 = \frac{C}{AB - C^2}$$

6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

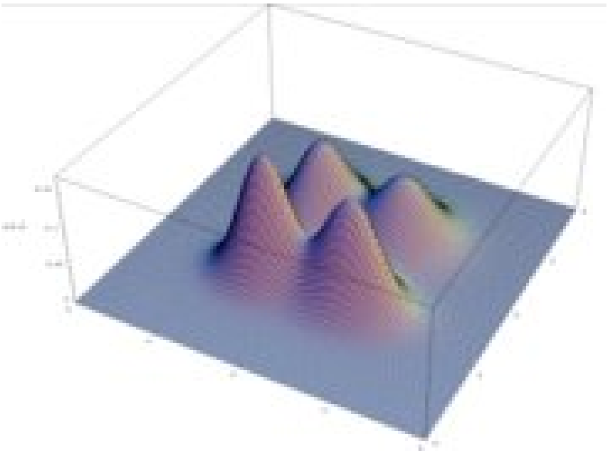
matriz de covariancia:

$$\begin{bmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{bmatrix} = \frac{1}{AB - C^2} \begin{bmatrix} -B & C \\ C & -A \end{bmatrix} = - \begin{bmatrix} A & C \\ C & B \end{bmatrix}^{-1}$$

- note que $AB - C^2$ é o determinante da matriz das derivadas
- a matriz de covariancia é o negativo do inverso da matriz de derivadas segundas
- $C = 0$ implica em $\sigma_{xy} = 0$: os parâmetros não são correlacionados
- $C = \pm\sqrt{AB}$: a matriz de derivadas segundas é singular (determinante $=0$); nesse caso X e Y estão relacionados linearmente

6. Inferência dos parâmetros de uma distribuição gaussiana bidimensional

- cuidado: $L(X, Y)$ pode ser multimodal!



7. Inferência dos parâmetros de uma distribuição gaussiana multidimensional

- vamos generalizar os resultados de uma distribuição gaussiana bidimensional para o caso multidimensional
- suponha que tenhamos um modelo com M parâmetros X_j, representados pelo vetor linha $\mathbf{X}$
- os valores de $\mathbf{X}$ que maximizam o posterior ($\mathbf{X_0}$) são obtidos resolvendo-se o sistema de M equações :

$$\left.\frac{\partial L}{\partial X_i}\right|_0 = 0$$

- vamos expandir o log do posterior em torno de $\mathbf{X_0}$ em uma série de Taylor até segunda ordem:

$$L = L(\mathbf{X_0}) + \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left.\frac{\partial^2 L}{\partial X_i \partial X_j}\right|_0 (X_i - X_{0i})(X_j - X_{0j}) + ...$$

7. Inferência dos parâmetros de uma distribuição gaussiana multidimensional

- expansão de Taylor até segunda ordem:

$$L = L(\mathbf{X_0}) + \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} \left.\frac{\partial^2 L}{\partial X_i \partial X_j}\right|_0 (X_i - X_{0i})(X_j - X_{0j}) + ...$$

- isso equivale a aproximar o posterior por uma gaussiana M-dimensional:

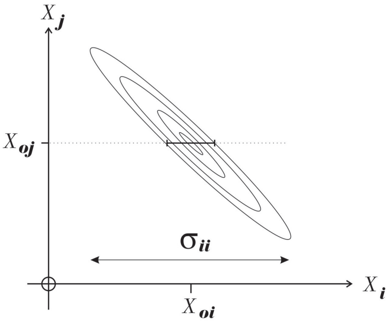
$$P(\mathbf{X}|D,H) \propto \exp\left[-\frac{1}{2}(\mathbf{X}-\mathbf{X_0})\mathbf{F}(\mathbf{X_0})\;(\mathbf{X}-\mathbf{X_0})^T\right]$$

- **F** é a matriz das derivadas segundas, com elementos $-\partial^2 L/\partial X_i \partial X_j$

7. Inferência dos parâmetros de uma distribuição gaussiana multidimensional

matriz de covariância:

$$[\sigma^2]_{ij} =< (X_i - X_{0i})(X_j - X_{0j}) >= -\mathbf{F}_{ij}^{-1}$$



7. Inferência dos parâmetros de uma distribuição gaussiana multidimensional

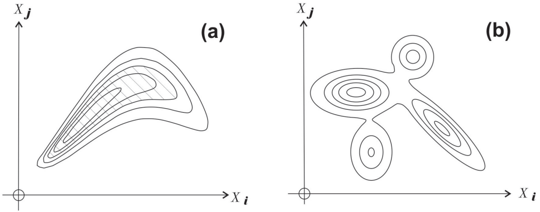


Figure: exemplos de distribuições onde a aproximação gaussiana não é boa

- Riess et al., 1998, ApJ, 116, 1009:
  Observational evidence from SNe for an accelerating universe and a cosmological constant
- modelo cosmológico ΛCDM
- dados $\{\mu_0\}$: conjunto dos módulos de distância para uma amostra de N SNs:

$$\mu = 5 \log D_l(z) + 25$$

  onde

$$D_l = D_l(z, \Omega_m, \Omega_\lambda, H_0)$$

- parâmetros do modelo: $\Omega_m, \Omega_\lambda, H_0$

8. Aplicação : a estatística de Riess et al. (1998)

Qual é o melhor modelo cosmológico que se pode inferir dos dados de SN?
Vamos usar o Teorema de Bayes:

- posterior dos parâmetros, dado os dados $\{\mu_0\}$ (esquecendo o H, das hipóteses a priori):

$$P(\Omega_m, \Omega_\lambda, H_0 | \{\mu_0\}) = \frac{P(\{\mu_0\} | \Omega_m, \Omega_\lambda, H_0) P(\Omega_m, \Omega_\lambda, H_0)}{P(\{\mu_0\})}$$

- como não temos vínculos a priori sobre os parâmetros, vamos supor que o prior $P(\Omega_m, \Omega_\lambda, H_0)$ é constante
- logo,

$$P(\Omega_m, \Omega_\lambda, H_0 | \{\mu_0\}) \propto P(\{\mu_0\} | \Omega_m, \Omega_\lambda, H_0)$$

- o melhor modelo, nesse caso, é o que maximiza a verossimilhança

verossimilhança do modelo: $P(\{\mu_0\}|\Omega_m, \Omega_\lambda, H_0)$

- supondo que as medidas são independentes:

$$P(\{\mu_0\}|\Omega_m, \Omega_\lambda, H_0) = \prod_{k=1}^{N} P(\mu_k|\Omega_m, \Omega_\lambda, H_0)$$

- supondo que os erros têm distribuição gaussiana:

$$P(\mu_i|\Omega_m, \Omega_\lambda, H_0) = \frac{1}{\sqrt{2\pi(\sigma_{\mu,i}^2 + \sigma_v^2)}} \times$$

$$\exp\left\{-\frac{[\mu(z_i, \Omega_m, \Omega_\lambda, H_0) - \mu_i]^2}{2(\sigma_{\mu,i}^2 + \sigma_v^2)}\right\}$$

onde $\sigma_{\mu,i}$ e σ_v são os erros no módulo de distância e no redshift (medido em unidades de módulo de distância)

8. Aplicação : a estatística de Riess et al. (1998)

verossimilhança do modelo: $P(\{\mu_0\}|\Omega_m, \Omega_\lambda, H_0)$

- logo,

$$P(\{\mu_0\}|\Omega_m, \Omega_\lambda, H_0) \propto \exp(-\frac{\chi^2}{2})$$

onde

$$\chi^2 = \sum_{i=1}^{N} \frac{[\mu(z_i, \Omega_m, \Omega_\lambda, H_0) - \mu_i]^2}{\sigma_{\mu,i}^2 + \sigma_v^2}$$

- o posterior normalizado será:

$$P(\Omega_m, \Omega_\lambda, H_0|\{\mu_0\}) = \frac{\exp(-\frac{\chi^2}{2})}{\int dH_0 \int d\Omega_\lambda \int \exp(-\frac{\chi^2}{2}) d\Omega_m}$$

- os parâmetros são obtidos maximizando essa expressão

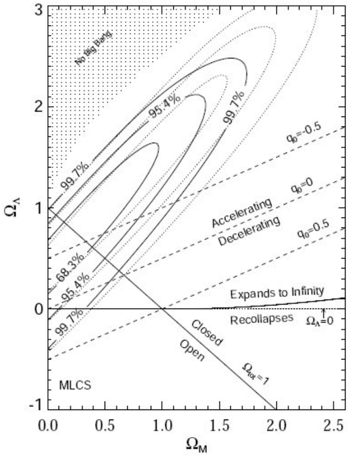


FIG. 6.—Joint confidence intervals for $(\Omega_M, \Omega_\Lambda)$ from SNe Ia. The solid contours are results from the MLCS method applied to well-observed SNe Ia light curves together with the snapshot method (Riess et al. 1998b) applied to incomplete SNe Ia light curves. The dotted contours are for the same objects excluding the unclassified SN 1997ck ($z = 0.97$). Regions representing specific cosmological scenarios are illustrated. Contours are closed by their intersection with the line $\Omega_M = 0$.

estimativa dos parâmetros

- pode-se examinar o comportamento da solução para Ω_m e Ω_λ, independentemente do valor de H_0, marginalizando sobre esse parâmetro:

$$P(\Omega_m, \Omega_\lambda | \{\mu_0\}) = \int P(\Omega_m, \Omega_\lambda, H_0 | \{\mu_0\}) dH_0$$

- pode-se estimar a probabilidade de Ω_λ ser positivo, marginalizando sobre Ω_m e só considerando os valores positivos de Ω_λ:

$$P(\Omega_\lambda > 0 | \{\mu_0\}) = \int_0^\infty d\Omega_\lambda \int P(\Omega_m, \Omega_\lambda | \{\mu_0\}) d\Omega_m$$

- ...

9. Comparação de modelos

- A história de A e B (Gull, 1989): *A tem uma teoria; B também tem uma teoria, mas com um parâmetro ajustável,* λ. *Que teoria devemos preferir com base nos dados* D*?*
- como comparar modelos de complexidade diferentes?
- podemos comparar as teorias de A e B pela razão de seus posteriores:

$$\frac{P(A|D,H)}{P(B|D,H)}$$

9. Comparação de modelos

- aplicando o teorema de Bayes, a razão dos posteriores fica:

$$\frac{P(A|D,H)}{P(B|D,H)} = \frac{P(D|A,H)}{P(D|B,H)} \times \frac{P(A|H)}{P(B|H)}$$

- para estimar seu posterior A precisa apenas dos dados, mas B, além dos dados, precisa de um valor para λ
- podemos comparar A e B se marginalizarmos sobre λ:

$$P(D|B,H) = \int P(D,\lambda|B,H)d\lambda = \int P(D|\lambda,B,H)P(\lambda|B,H)d\lambda$$

- vamos supor também que, a priori B pode supor que λ tem uma distribuição uniforme entre λ_{min} eλ_{max} :

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

para λ entre λ_{min} e λ_{max} e 0 fora desse intervalo

- vamos supor que tem um valor λ_0 que provê o melhor ajuste às medidas, e vamos então aproximar a verossimilhança dos dados para o modelo B por uma gaussiana de largura σ_λ:

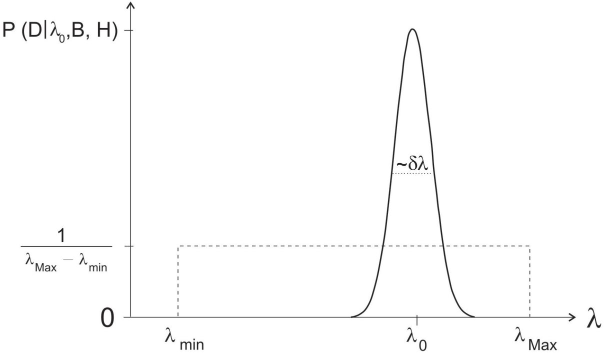
$$P(D|\lambda, B, H) = P(D|\lambda_0, B, H) \exp\left[-\frac{(\lambda - \lambda_0)^2}{2\sigma_\lambda^2}\right]$$

- então,

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

$$= \frac{P(D|\lambda_0, B, H)\sigma_\lambda\sqrt{2\pi}}{\lambda_{max} - \lambda_{min}}$$

9. Comparação de modelos



- a razão dos posteriores então fica:

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

- o primeiro termo reflete nossas preferências relativas a priori sobre as teorias: para não ter-se viés, é razoável supor que ele vale 1
- o segundo termo é a razão entre as melhores previsões de cada modelo para os dados disponíveis; é de se esperar que a verossimilhança de B seja maior que a de A, devido ao parâmetro adicional
- o terceiro termo é o fator de Occam: penaliza o modelo B por seu parâmetro adicional
- a navalha de Occam = princípio da simplicidade:
  Frustra fit per plura quod potest fieri per pauciora
  é bobagem fazer com mais o que pode se fazer com menos (sec.XIV)

10. Dark Energy Task Force

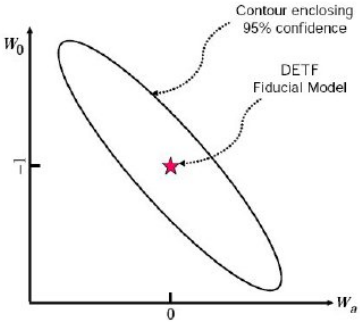
- relatório da DETF: Albrecht et al. (astro-ph/0609591)
- objetivos:
  - estudar as estratégias para abordar o problema da Energia Escura (DE)
  - analisar as propostas atuais e futuras, com base numa figura de mérito bem definida
  - fazer recomendações
  - assinado por Rocky Kolb, do Fermilab

10. Dark Energy Task Force

- parametrização da DE: $p = w\rho$

$$w(a) = w_0 + (1 - a)w_a$$

- w_0: valor atual de w
- w_a: parametriza a evolução de w
- constante cosmológica: $w_0 = -1$, $w_a = 0$
- figura de mérito: inverso da área contida pelo limite de confiança de 95% no plano $w_0 \times w_a$


Contour enclosing
95% confidence
DETF
Fiducial Model
W_0
−1
0
W_a

10. Dark Energy Task Force

- "sensores" de DE: a expansão $[H(z)]$ e o crescimento de estruturas $[f(z) = d \ln \rho / d \ln t]$
- observáveis:
  - SN Ia: $H(z)$
  - BAOs: $H(z)$ (os mesmos dados são sensíveis a f se o bias b e o parâmetro de distorção β são conhecidos)
  - tomografia com lentes fracas: $H(z)$ e f
  - estatística de aglomerados: $H(z)$ e f

10. Dark Energy Task Force

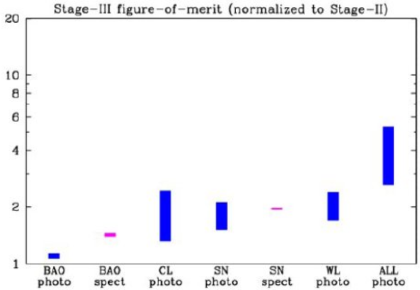


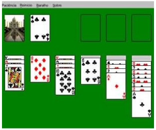
Figure: FoM para várias técnicas fotométricas (photo) e espectroscópicas (spec). A extensão da barra depende dos efeitos sistemáticos- se pessimistas ou otimistas. *All photo* refere-se á combinação dos resultados fotométricos de BAO,CL,SN e WL.

11. MCMC: Markov Chain Monte Carlo
I. Motivação

- amostrar uma distribuição de probabilidades arbitrária
- eficiente para amostrar distribuições complexas
- permite resolver problemas difíceis
- computacionalmente intensiva

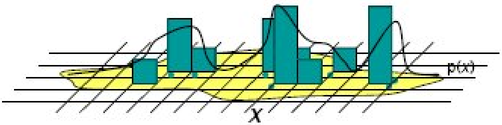
11. MCMC: Markov Chain Monte Carlo
II. Princípios de Monte Carlo

- qual é a probabilidade de se ganhar uma partida de *Paciência* com cartas bem embaralhadas?
- problema difícil de ser abordado analiticamente porque ganhar ou perder depende de um procedimento complexo de organizar as cartas
- solução: jogue muitas vezes e estime a probabilidade de ganhar empiricamente
- de modo geral: aproxime uma função de distribuição de probabilidades por um conjunto de amostras extraídas dessa distribuição



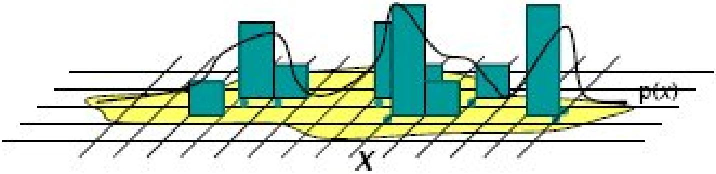
11. MCMC: Markov Chain Monte Carlo
II. Princípios de Monte Carlo

- é dada uma distribuição de probabilidades $P(x)$ (que pode ser de muitas variáveis)
- extrai-se dessa distribuição N amostras independentes e identicamente distribuídas (i.i.d.):
  todas as amostras têm a mesma distribuição de probabilidades e são independentes umas das outras
- podemos fazer uma aproximação de $P(x)$ com essas amostras:



11. MCMC: Markov Chain Monte Carlo
II. Princípios de Monte Carlo

- N amostras i.i.d.:



$$\mathrm{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N} 1(x^{(i)} = x) \underset{N\to\infty}{\longrightarrow} \mathrm{p}(x)$$

- aproximação de $P(x)$: histograma n-dimensional

$$P'(x) \simeq \frac{1}{N}\sum_{i=1}^{N} 1(\text{se } x_i = x) \longrightarrow P(x) \text{ para } N \to \infty$$

- estimativa de valores esperados de funções de x:

$$< f > \simeq \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

- estimativa do máximo de $P(x)$:

$$x^{max} \simeq \mathrm{argmax}_{x_i}[P(x_i)]$$

- suponha que temos uma variável x distribuída uniformemente entre 0 e 1:

$$P(x)dx = \begin{cases} dx & 0 < x < 1 \\ 0 & \text{nos outros casos} \end{cases}$$

- suponha que tenhamos outra variável, $y = y(x)$ e que desejemos $P(y)dy$
- lei das transformações de probabilidades:

$$|P(x)dx| = |P(y)dy|$$

- exemplo: distribuição exponencial para $0 < y < \infty$

$$P(y|a) = \frac{1}{a} e^{-y/a}$$

- logo, como $|P(x)dx| = |P(y)dy|$,

$$\gamma = \int_0^{\gamma} P(x)dx = \int_0^{y} P(y|a)dy = 1 - e^{-y/a}$$

  ou

$$y = -a \ln(1 - \gamma)$$

- logo, para se gerar N pontos y_i com uma distribuição $P(y|a)$, gera-se N números aleatórios γ_i, distribuídos uniformemente entre 0 e 1 e calcula-se $y_i = -a \ln(1 - \gamma_i)$
- infelizmente, a integração direta só é possível para distribuições de probabilidades muito simples

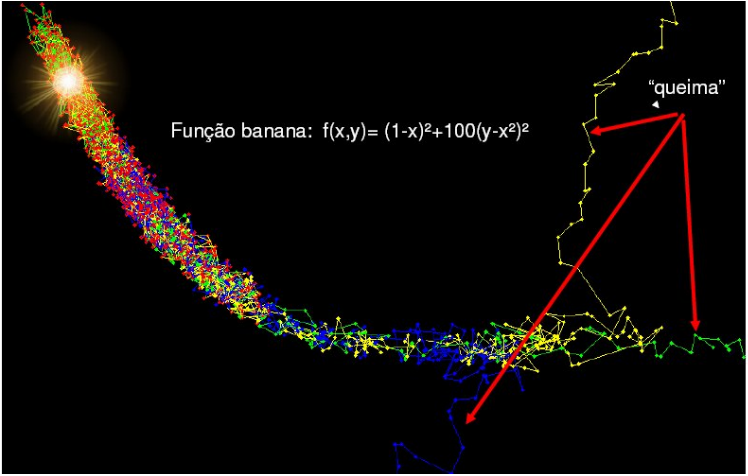
- objetivo do MCMC: gerar uma sequência de pontos no espaço de parâmetros cuja distribuição é $P(x)$
- processo de Markov: a próxima amostra depende da atual mas não das anteriores
- a geração de novas amostras deve obedecer o princípio do "balanceamento detalhado":
  a probabilidade de se ir de um ponto x_1 no espaço de parâmetros a um ponto x_2 é igual à probabilidade de se ir de x_2 a x_1

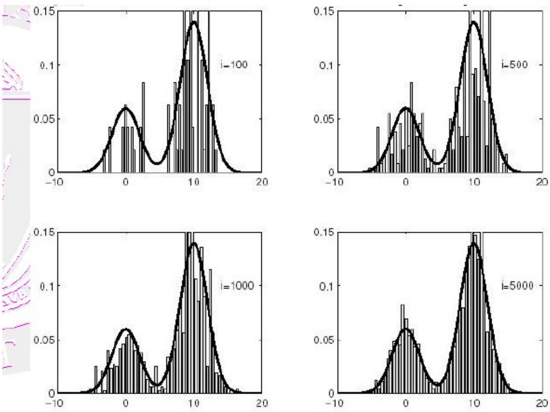
- algoritmo de Metropolis:
  - comece num ponto x_1 e determine um novo ponto x_2:

$$x_2 = x_1 + \epsilon$$

    (ϵ pode ser sorteado com uma distribuição gaussiana)
  - se $P(x_2) > P(x_1)$ aceita o novo ponto
  - se $P(x_2) < P(x_1)$ aceita com probabilidade $P(x_2)/P(x_1)$
  - se um novo ponto é aceito, faz-se $x_1 = x_2$
  - determina-se um novo x_2 e repete-se o processo
- se $P(x) \propto \exp(-E(x))$, então
  - se $E(x_2) < E(x_1)$ aceita
  - se $E(x_2) > E(x_1)$ aceita com probabilidade $\exp[-(E(x_2) - E(x_1))]$

11. MCMC: Markov Chain Monte Carlo
III. Cadeias de Markov





$q(x^*|x) \sim N(x^{(i)}, 100)$

$p(x) \sim 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x-10)^2)$

III. Cadeias de Markov

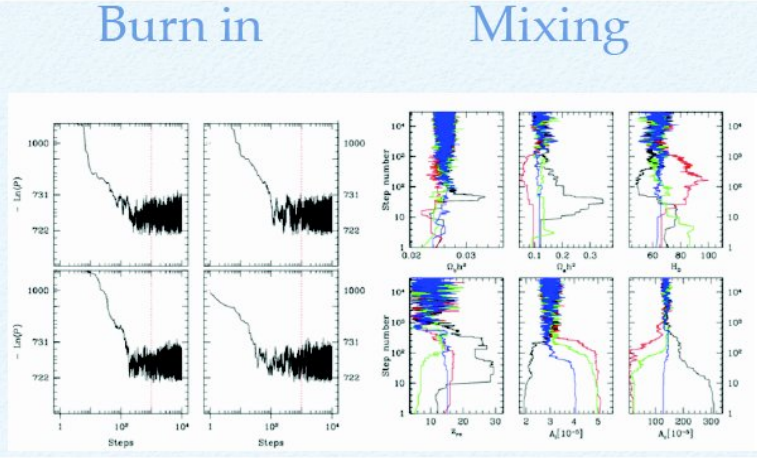
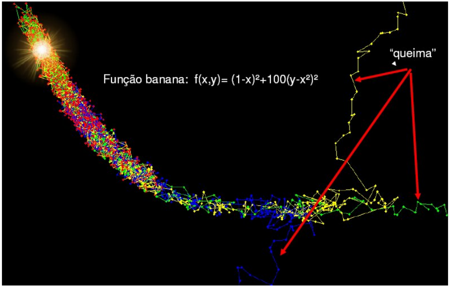


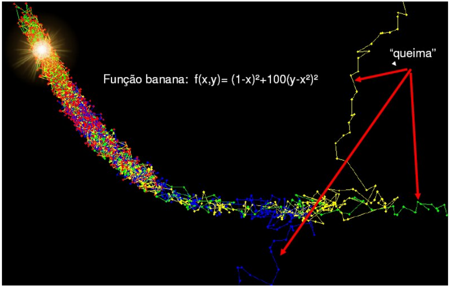
Figure: Diagnóstico de MCMC.

- na prática:
  - trabalha-se com várias cadeias (>4)
  - desprezam-se as primeiras amostras de cada cadeia ("queima")
  - aplicam-se testes para verificar se as cadeias convergiram



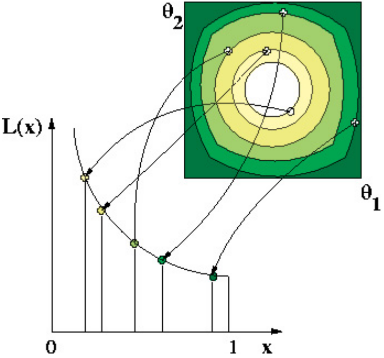
11. MCMC: Markov Chain Monte Carlo
III. Cadeias de Markov

- obtendo-se amostras do posterior via MCMC pode-se resumir a inferência via:
  - melhor solução, média, mediana, moda
  - regiões de confiança ("erros")
  - resultados sem os *nuisance parameters*



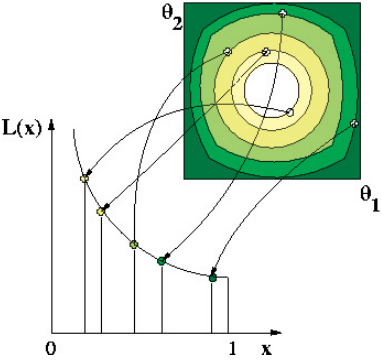
12. Nested Sampling

- o melhor jeito de calcular a evidência
- amostra o posterior melhor que MCMC!
- inventado por John Skylling em 2005
- (ver Sivia & Skilling)

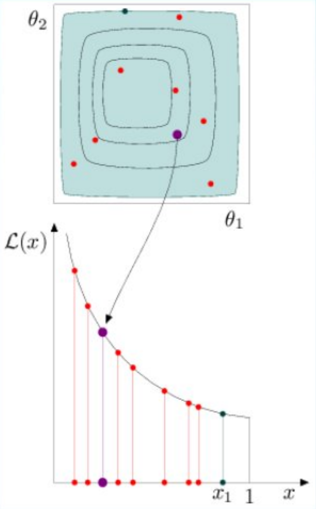


12. Nested Sampling

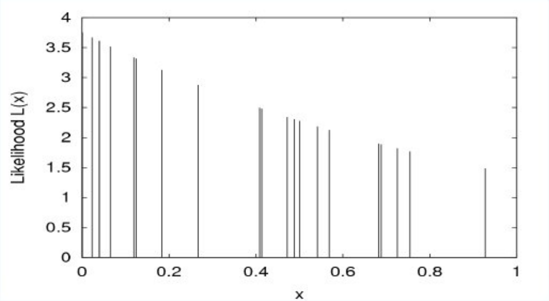
- evidência: $E = \int p(D|\theta, H)p(\theta|H)d\theta$
- ideia básica: mudança de variável: $dX = p(\theta|H)d\theta$
- sendo $L = p(D|\theta, H)$, vem que $E = \int_0^1 L dX$



- Imagine sorting all points in parameter space by likelihood
- Then map the parameter space onto [0,1] in order of decreasing likelihood
- If we could get some points then we can do the integral numerically, easily!
- Need more points with small x than large x



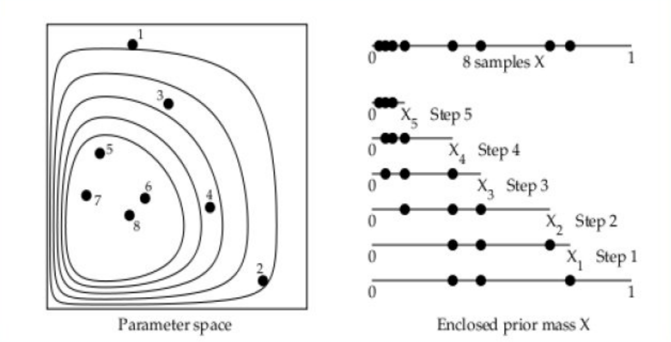
- The hard thing is getting the x value. i.e. Here's a point, what fraction of the prior probability lies at a higher likelihood?
- But we can *estimate* the x's.



The Algorithm

- Generate n points from the prior
- Loop for i=1,2,3,...
- {
- Find the one with the lowest likelihood, L_{worst}. Remove it from the population, but store it for results. Estimate its x value as $(n/(n+1))^i$.
- Replace that point with a new one generated from the prior, but subject to the hard constraint $L>L_{worst}$.
- }

- Slight modification necessary in case of a tie for the worst point. I won't go into detail about this here.



How to Generate That New Point

- This part of the algorithm was swept under the rug in the previous slide:
- Replace that point with a new one generated from the prior, but subject to the hard constraint $L > L_{worst}$.
- Generally quite hard to do this
- An easy but approximate way: Copy a surviving point (guaranteed to have $L > L_{worst}$) and evolve it with MCMC according to the prior, but rejecting if $L <= L_{worst}$.

1. Veja o que diz o *Numerical Recipes* sobre métodos bayesianos e procurem no Google por *bayesian versus frequentist*.
2. Em um jogo há 3 caixas (A, B e C) e dentro de uma delas se esconde uma prenda, que se leva quando se acerta a caixa. Você escolhe uma caixa (digamos a A) e, a seguir, o apresentador do jogo abre uma das outras duas para mostrar que a prenda não está nela. Nesse ponto você pode tanto manter sua escolha original (caixa A) quanto mudar para outra (no caso B). Para ganhar a prenda é melhor manter ou mudar a escolha original? (Trotta)
3. Reflita sobre o fator de Occam. Se o intervalo do parâmetro adicional for infinito, haverá uma penalização infinita a ele. Isso é razoável?

Exercícios

4. Um dos testes propostos por Einstein para a Teoria da Relatividade Geral foi a deflexão da luz das estrelas pelo Sol, que ele previu como $\alpha = 4GM/(c^2R) = 1.74$ arcsec, onde M e R são a massa e o raio do Sol. Em 1919 duas expedições , uma em Sobral e outra na Ilha do Príncipe, mediram α obtendo, respectivamente, 1.98 ± 0.16 arcsec (baseado em 7 estrelas) e 1.61 ± 0.40 (baseado em 5 estrelas). Qual é o fator de Bayes entre a gravitação de Einstein e a newtoniana a partir desses dados? Comente sobre a força da evidência. (Trotta)

Projeto: SN Ia

Introdução:

- vamos fazer estimativas bayesianas de dados de SN Ia
- adaptação de um exercício proposto por Roberto Trotta na *I Jayme Tiomno School of Cosmology*
- antes reveja a seção 8: artigo de Riess et al. (1998)

Projeto: SN Ia

- vamos supor que temos um conjunto de dados de SN Ia: SNe_simulated.txt: 300 valores de redshift z_i e magnitude m_i no máximo
- o módulo de distância de cada SN é

$$\mu = m - M = 5 \log D_l + 25$$

  onde a distância de luminosidade D_l é em Mpc
- D_l depende da cosmologia via

$$E(z) = [\Omega_m (1+z)^3 + \Omega_\lambda + \Omega_k (1+z)^2]^{1/2}$$

$$\Omega_k = 1 - \Omega_m - \Omega_\lambda$$

- vamos supor que $h = 0.72$ e que os parâmetros do modelo são $\theta = (\Omega_m, \Omega_\lambda)$

Projeto: SN Ia

- vamos supor que os módulos de distância $\mu_i = m_i - M$ tem erro gaussiano, σ_i, e que sua verossimilhança é

$$P(\mu_i|M,\theta,\sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i^2} \exp\left[-\frac{(\mu_i - \mu(z_i,\theta))^2}{2\sigma_i^2}\right]$$

  vamos supor que $\sigma_i = \sigma = 0.4$
- vamos supor, inicialmente, que $M = M_0 = -19.3$
- Exercício 1: amostre o posterior (mcmc, NS), estime os valores mais prováveis de $\theta = (\Omega_m, \Omega_\lambda)$ e plote as regiões com 68% e 95% de probabilidade no plano $\Omega_m \times \Omega_\lambda$. Deixe clara sua escolha de priores.
- Exercício 2: idem, mas supondo $\Omega_k = 0$.

Projeto: SN Ia

- Exercício 3: volte ao exercício 1 mas suponha agora que M_0 tem uma distribuição gaussiana: $N(-19.3, \sigma)$
- Exercício 4: suponha que σ também não é conhecido. Como se trata de um parâmetro de escala, recomenda-se que se use um prior do tipo $p(\sigma) \propto 1/\sigma$.
- Exercício 5: a localização dos picos no espectro de potência se traduz num vínculo entre Ω_m e Ω_λ:

  $$1.41\Omega_\lambda + \Omega_m = 1.30 \pm 0.04$$

  Adicione este vínculo, supondo uma distribuição gaussiana.
- Exercício 6: podemos, adicionalmente, incluir o vínculo obtido pela observação de BAOs: $D_A(z = 0.3) = (893 \pm 27)$ Mpc Adicione este vínculo, supondo uma distribuição gaussiana. Lembre-se que $D_A(z) = (1+z)^2 D_l(z)$.
- Aproximando o posterior unidimensional $p(\Omega_\lambda|...)$ por uma gaussiana, estime o fator de Bayes entre um modelo com $\Omega_\lambda = 0$ e ΛCDM.

Referências

- referência boa: Sivia e Skilling, *Data Analysis - A Bayesian Tutorial*
- arXiv:0906.0664: Statistical techniques in cosmology, A. Heavens
- School of Astrophysics "Francesco Lucchin": **http://www.bo.astro.it/~school09/prog.html**
- III INPE Advanced Course on Astrophysics: Astrostatistics **www.das.inpe.br/school/2009/lectures.html**
- Statistical methods for Cosmology, Roberto Trotta: ver notas de aula da *I Jayme Tiomno School of Cosmology*: **http://www.on.br/jaymeschool/school_presentations.html**
- aplicações de MCMC: Verde et al. 2003, ApJS, 148, 195

Referências

- D. MacKay, *Information Theory, Inference, and Learning Algorithms*: http://www.inference.phy.cam.ac.uk/mackay/
- E. T. Jaynes, *Probability Theory: The Logic of Science* http://bayes.wustl.edu/etj/prob/book.pdf
- Mukherjee, Pia, et al. (2006), ApJ, 638,L51 *A Nested Sampling Algorithm for Cosmological Model Selection*

Pacotes públicos de cosmologia bayesiana

- CosmoMC (Sarah Bridle & Antony Lewis)
  http://cosmologist.info/cosmomc/
- SuperBayes (Ruiz de Austri, Feroz & Trotta)
  http://www.ft.uam.es/personal/rruiz/superbayes/index.php?page=m
- CosmoNest (Parkinson, Mukherjee & Liddle)
  http://cosmonest.org/