Inferência para CS
Seleção de Modelos

Renato Martins Assunção

DCC - UFMG

2025

Entropy

- Imagine a sequence of independent symbol emissions from a source.
- Symbols X are selected randomly from a finite dictionary according to a probability distribution $p(x)$.
- Each symbol is represented with a 0-1 string
- Prefix code is used:
  - No codeword appears at the beginning (prefix) of any other codeword.
  - Avoid ambiguity when decoding
  - Instantaneous decoding: no need to wait to see the next symbol.
- We want to use the shortest possible code to save on transmission.
- Can you find this shortest code?

Entropy

- Yes, we can.
- Shannon's Mater's thesis (1937)
- For a discrete random variable X, the Shannon entropy is defined as:

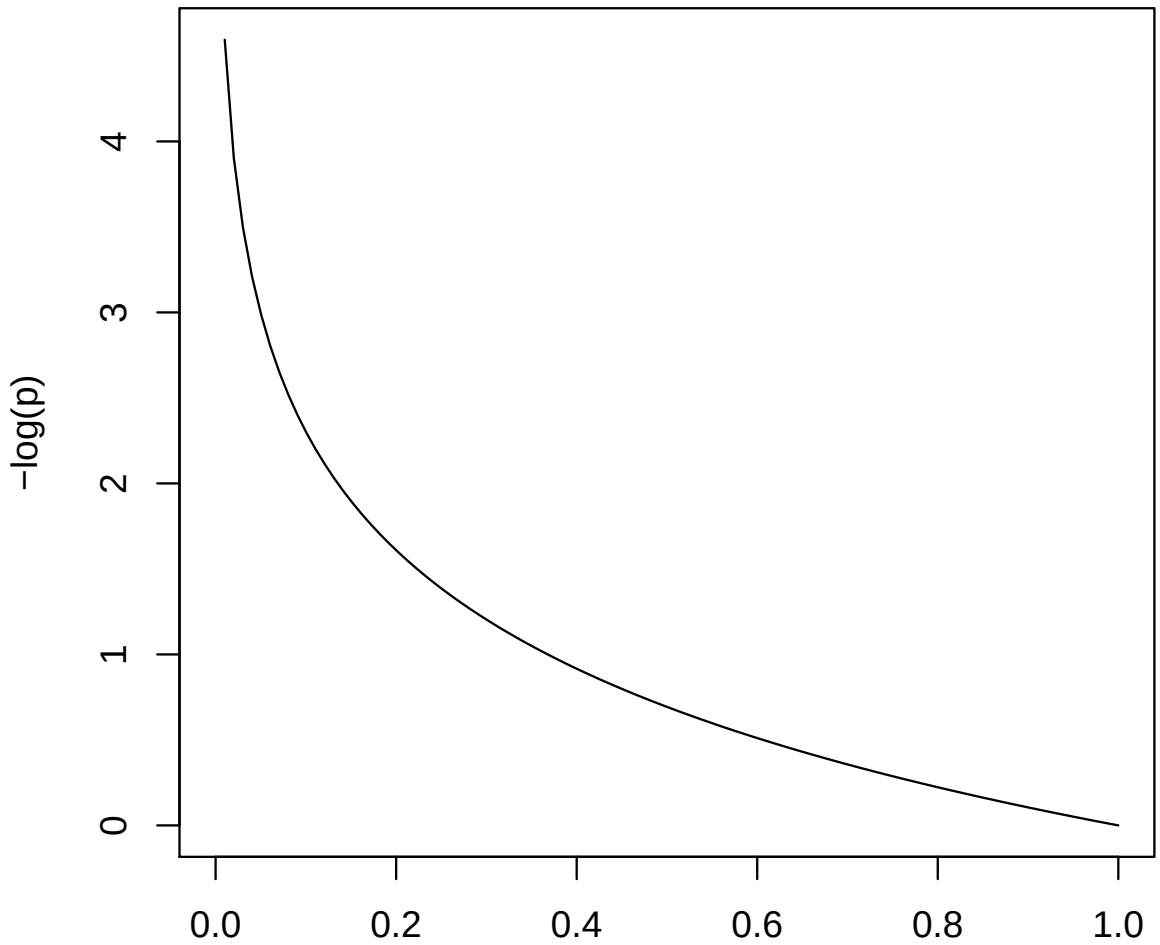
$$H(X) = -\sum_{x} p(x) \log_2(p(x))$$

- The entropy of a probability distribution is the minimum expected number of bits required to encode symbols drawn from that distribution.
- It is not the minimum number of bits needed for each individual symbol.
- It is the minimum average (expected) number of bits per symbol, over many symbols, when using the best possible encoding strategy.
- We will adopt a different approach

Entropia de um evento

- Seja E um evento num certo espaço amostral.
- O evento E ocorre ou não ocorre em cada realização do experimento aleatório.
- Seja $\mathbb{P}(E) \in [0, 1]$ a probabilidade da ocorrência de E.
- A entropia associada com a ocorrência do evento E mede o GRAU DE SURPRESA que a ocorrência de E acarreta.
- Surprise: $-log(\mathbb{P}(E))$.
- Why???? In just few minutes...

Entropia de evento é $-\log(p)$



Log em que base?

- Qualquer uma.
- Como $\log_a(x) = \log_a(b) \ \ \log_b(x)$ temos

$$\log_a(x) = c \ \log_b(x)$$

onde $c = \log_a(b)$ é uma constante que depende apenas das duas bases, e não de x.

- Isto implica que a diferença absoluta de log's numa base a é igual á diferença na base b vezes uma constante

$$\log_a(x) - \log_a(y) = c \ (\log_b(x) - \log_b(y))$$

- E diferenças relativas são iguais nas duas bases

$$\frac{\log_a(x)}{\log_a(y)} = \frac{\log_b(x)}{\log_b(y)}$$

- A idéia é que muita ou pouca entropia numa base será também muita ou pouca entropia na outra base.

Interpretação de entropia

- Seja a um inteiro entre 1 e 9. Temos

$$0 = \log_{10}(1) \leq \log_{10}(a) \leq \log_{10}(9) \approx 0.95$$

- Seja $p = a.bcdef...10^{-k}$ onde a é um inteiro entre 1 e 9.
- Tome entropia na base 10. Então

$$\begin{aligned} -\log_{10}(p) &= -\log_{10}(a.bcdef \ldots 10^{-k}) \\ &= -\log_{10}(a.bcdef \ldots) - \log_{10}(10^{-k}) \\ &= -\log_{10}(a.bcdef \ldots) + k \\ &\approx k \end{aligned}$$

já que o primeiro termo é um valor entre 0 e -1 (ver 1a equação).

- Então: ENTROPIA de p é $-\log_{10}(p)$ (aprox) o número de casas decimais antes do primeiro número significativo.

Interpretação de entropia

- Se tomarmos entropia com logs na base 2 (isto é, entropia é $-\log_2(p)$), então a entropia será aprox o número de bits iguais a zero na expansão de p na base 2 antes do primeiro bit significativo.
- Um indivíduo escolhe um número entre $\{0, 1, 2, \ldots, 9\}$
- Uma loteria sorteia um dos números da lista com igual probabilidade.
- A chance de acertar na loteria é $p = 0.1$ com entropia (surpresa) $-\log_{10}(p) = 1$.
- Suponha agora que a lista de números seja $\{0, 1, 2, \ldots, 99\}$.
- A entropia do evento acertar na nova loteria (surpresa) é $-\log_{10}(0.01) = 2$.
- Se a lista de números for $\{0, 1, 2, \ldots, 999\}$.
- a entropia (surpresa) passa a ser $-\log_{10}(0.001) = 3$.

Interpretação de entropia

- Incremento na surpresa é linear com diminuição multiplicativa da probabilidade.
- Incremento de surpresa de ganhar na loteria quando passo de probabilidade 0.1 para 0.01 é $\Delta = 2 - 1 = 1$.
- Incremento ao passar 0.01 para 0.001 é TAMBÉM $\Delta = 3 - 2 = 1$.
- De maneira geral:

$$-\log_{10}\left(\frac{p}{10^k}\right) = -\log_{10}(p) + k$$

- Dividir por 10^k a chance faz aumentar a surpresa em k unidades.
- Surpresa (entropia) cresce linearmente com a ORDEM DE GRANDEZA (ou precisão) de p.

Entropia tem de ser da forma log

- Se entropia (surpresa) funciona desta forma, ela TEM DE SER da forma $-\log(p)$.
- Por quê?
- O que significa "funcionar desta forma"??
- Seja $S : [0, 1] \to [0, \infty)$ uma função matemática que visa capturar o sentido de surpresa.
- Queremos que S tenha as seguintes propriedades óbvias:
  1. $S(1) = 0$ (a ocorrência de um evento que tem chance 100% de ocorrência tem surpresa 0, nula).
  2. $S(0) = \infty$ (a ocorrência de um evento impossível traz surpresa infinita)
  3. $S(p)$ é decrescente em p

Propriedade adicional

- Vamos impor uma condição adicional em $S(p)$.
- A função $S(p)$ deverá satisfazer a seguinte propriedade:

$$S(p_2 p_1) - S(p_2) = S(p_3 p_1) - S(p_3)$$

  para todo p_1, p_2, p_3 em $[0, 1]$.
- O que esta propriedade está dizendo?
- Tome o aumento de surpresa $S(p_2 p_1) - S(p_2)$ ao passar da ocorrência de um evento com probabilidade p_2 para outro com probabilidade menor $p_2 p_1$.
- Este aumento de supresa é o mesmo se reduzimos a probabilidade p_3 de um evento por p_1 passando então a ter a probabilidade $p_3 p_1$.

Propriedade adicional

- Suponha
$$S(p_2 p_1) - S(p_2) = S(p_3 p_1) - S(p_3)$$
para todo p_1, p_2, p_3 em $[0, 1]$.
- Por exemplo, ao passar de $p_2 = 0.5$ para $p_2 p_1 = 0.5/5 = 0.1$ teremos certo aumento Δ de surpresa.
- Este aumento Δ de surpresa é o mesmo que temos ao passar de $p_3 = 0.0003$ para $p_3 p_1 = 0.0003/5 = 0.00006$.
- E isto vale para todo p_1, p_2, p_3.
- Este é o sentido desta propriedade adicional que queremos para a função surpresa.

Teorema

- Uma função S com estas 4 propriedades só pode ser da forma $S(p) = -c\log(p)$, onde c é uma constante positiva qualquer.
- **PROVA:** Tome $p_3 = 1$ na propriedade adicional.
- Como $S(1) = 0$, temos

$$S(p_2 p_1) - S(p_2) = S(p_3 p_1) - S(p_3) = S(p_1) - S(1) = S(p_1)$$

- e então

$$S(p_2 p_1) = S(p_2) + S(p_1)$$

- Isto é, a função $S(p)$ deve transformar produtos em somas.
- A única função com esta propriedade é a função log
- Ver prova disso em livros de análise matemática.

Surpresa acarretada pela ocorrência de v.a. X

- Suponha que X seja uma v.a. discreta com a seguinte distribuição:

Valores Possíveis	x_1	x_2	$\ldots$	x_M
Probabilidades	$p(x_1)$	$p(x_2)$	$\ldots$	$p(x_M)$
Surpresas	$-\log p(x_1)$	$-\log p(x_2)$	$\ldots$	$-\log p(x_M)$

- Um valor aleatório de X é selecionado com as probabilidades acima.
- Suponha que o valor instanciado de X seja x_i
- Se o valor x_i for raro, a surpresa $-\log p(x_i)$ ocasionada por sua ocorrência será grande.
- Se o valor x_i for comum, a surpresa $-\log p(x_i)$ será pequena.
- A surpresa é uma variável aleatória: $-\log p(X)$.
- Qual a surpresa ESPERADA se repetirmos o procedimento de selecionar X com a distribuição acima?

Entropia de v.a. discreta

- Qual a surpresa ESPERADA que a quantidade aleatória $-\log p(X)$ acarreta?

Valores Possíveis	x_1	x_2	$\ldots$	x_M
Probabilidades	$p(x_1)$	$p(x_2)$	$\ldots$	$p(x_M)$
Surpresas	$-\log p(x_1)$	$-\log p(x_2)$	$\ldots$	$-\log p(x_M)$

- Esperança é a soma de cada valor possível vezes sua probabilidade de ocorrência

$$\begin{aligned} \mathbb{H}_p(p) &= \sum_{i=1}^{n} -\log(p(x_i))\ p(x_i) \\ &= \mathbb{E}_p\left\{-\log\left(p(X)\right)\right\} \end{aligned}$$

- Esta fórmula é a definição de entropia de uma distribuição de probabilidade discreta.

Entropia de v.a. discreta

- Entropia de v.a. discreta:

$$\mathbb{H}_p(p) = \sum_{i=1}^{n} -\log(p(x_i))\ p(x_i) = \mathbb{E}_p\left\{-\log\left(p(X)\right)\right\}$$

- Estude esta última notação.
- Perceba que $p(X)$ é o valor de $p(x_i)$ na tabela escolhido ao acaso como função do valor de X.
- A variável X, por sua vez, é selecionada com as mesmas probabilidades p da tabela.
- O sub-índice p sob o símbolo da função esperança e em $\mathbb{H}_p(p)$ é para enfatizar que o X aleatório no argumento da função possui distrbuição dada por p na tabela acima.
- A notação $\mathbb{H}_p(p)$ parece redundante mas ela será útil quando definirmos a distância de Kullback-Leibler.

Exemplo

- X é v.a. discreta com M valores equiprováveis.
- Isto é, $\mathbb{P}(X = x_i) = 1/M$ para todo valor x_i, com $i = 1, 2, \ldots, M$.
- Então $\mathbb{H}_p(p)$ é dada por

$$\begin{aligned} \mathbb{E}_p \{-\log(p(X))\} &= \sum_{i=1}^{M} -\log\left(\frac{1}{M}\right)\frac{1}{M} \\ &= M\frac{1}{M}\left(-\log M^{-1}\right) \\ &= \log(M) \end{aligned}$$

- Assim, a entropia $\mathbb{H}_p(p)$ de uma uniforme é o logaritmo do número de classes equiprováveis.

Exemplo

- X é v.a. com distribuição Poisson com parâmetro λ.
- Então

$$\begin{aligned}\mathbb{E}_p\{-\log(p(X))\} &= \sum_{i=0}^{\infty} -\log\left(\frac{\lambda^k e^{-\lambda}}{k!}\right)\frac{\lambda^k e^{-\lambda}}{k!} \\ &= \lambda - \sum_{i=0}^{\infty}\log\left(\frac{\lambda^k}{k!}\right)\frac{\lambda^k}{k!}e^{-\lambda}\end{aligned}$$

- A entropia $\mathbb{H}_p(p)$ da distribuição de Poisson não possui uma expressão mais simples que esta.

Caso contínuo

- Se X é uma v.a. contínua com densidade $f(x)$ então

$$\mathbb{H}_f(f) = \int_{\mathbb{R}} -\log(f(x))\ f(x)\ dx = \mathbb{E}_f\left[-\log f(X)\right]$$

  onde o sub-índice f na esperança indica que X é selecionada com densidade f.
- Podemos pensar num procedimento em três etapas:
  - Tome $X \sim f$
  - Tome a altura aleatória $f(X)$ da densidade.
  - Tome a esperança de $-\log(f(X))$.

Exemplo - normal

- Suponha que $X \sim N(\mu, \sigma^2)$.
- Neste exemplo, ao invés de integramos uma função, podemos usar o fato de que a variável aleatória padronizada $Z = (X - \mu)/\sigma$ possui distribuição $N(0, 1)$.
- Portanto, $\mathbb{E}(Z) = 0$ e $\mathbb{V}(Z) = \mathbb{E}(Z^2) = 1$.
- Temos

$$\begin{aligned} -\log f(x) &= -\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\right] \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \end{aligned}$$

Exemplo - normal

- Portanto

$$\begin{aligned}
\mathbb{H}_f(f) &= -\mathbb{E}_f\left[\log f(X)\right] \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\mathbb{E}_f\left(\frac{X-\mu}{\sigma}\right)^2 \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} \\
&= \frac{1}{2}\log(2\pi e\sigma^2)
\end{aligned}$$

Exemplo - normal

- Se $X \sim N(\mu, \sigma^2)$ então $\mathbb{H}_f(f) = 0.5 \log(2\pi e \sigma^2)$.
- Veja que a entropia depende apenas de σ e não de μ.
- Além disso, a entropia aumenta com σ numa escala logarítimica.
- Outro fato curioso: com v.a.'s contínuas, a entropia pode ser negativa se $\sigma^2 < 1/(2\pi e)$.
- As vezes, vamos escrever simplemenste $\mathbb{H}(X)$ para significar $\mathbb{H}_f(f)$

Teoria da informação

- Minimum Description Length:
- Entropy represents the minimum average number of bits needed to encode the outcomes of the random variable.
- On average, no more efficient code can exist to represent the information in the random variable.

Medindo distâncias entre distribuições

- Ao comparar duas grandezas físicas A e B (tais como duas massas, duas velocidades ou duas cargas elétricas) sabemos dizer se A e B são aproximadamente iguais ou muito diferentes.
- E ao comparar as *distribuições* de duas variáveis aleatórias X e Y, quando podemos dizer que elas são muito diferentes? Como medir isto?
- Exemplo: $X \sim N(0,1)$, $Y \sim N(1,1)$ e $Z \sim N(5,1)$,
- Esperamos que a distribuição de Y seja próxima daquela de X e afastada da distribuição de Z.
- Mas e se $Z \sim N(0.5, 2^2)$? Qual seria mais próxima de X? Y ou Z?

Comparando gaussianas

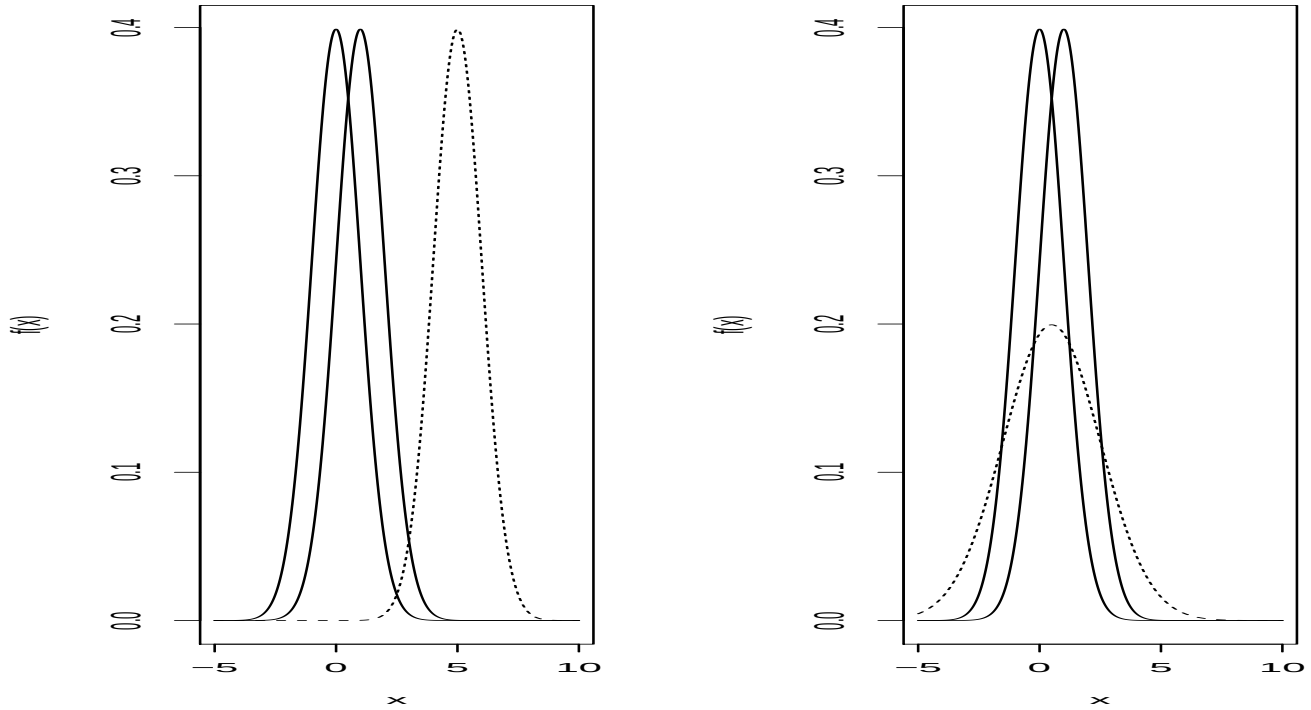


Figura: Esquerda: $N(0,1)$ e $N(1,1)$ em linha sólida e $N(5,1)$ em linha tracejada. Direita: $N(0,1)$ e $N(1,1)$ em linha sólida e $N(0.5,2^2)$ em linha tracejada.

Comparando densidades arbitrárias

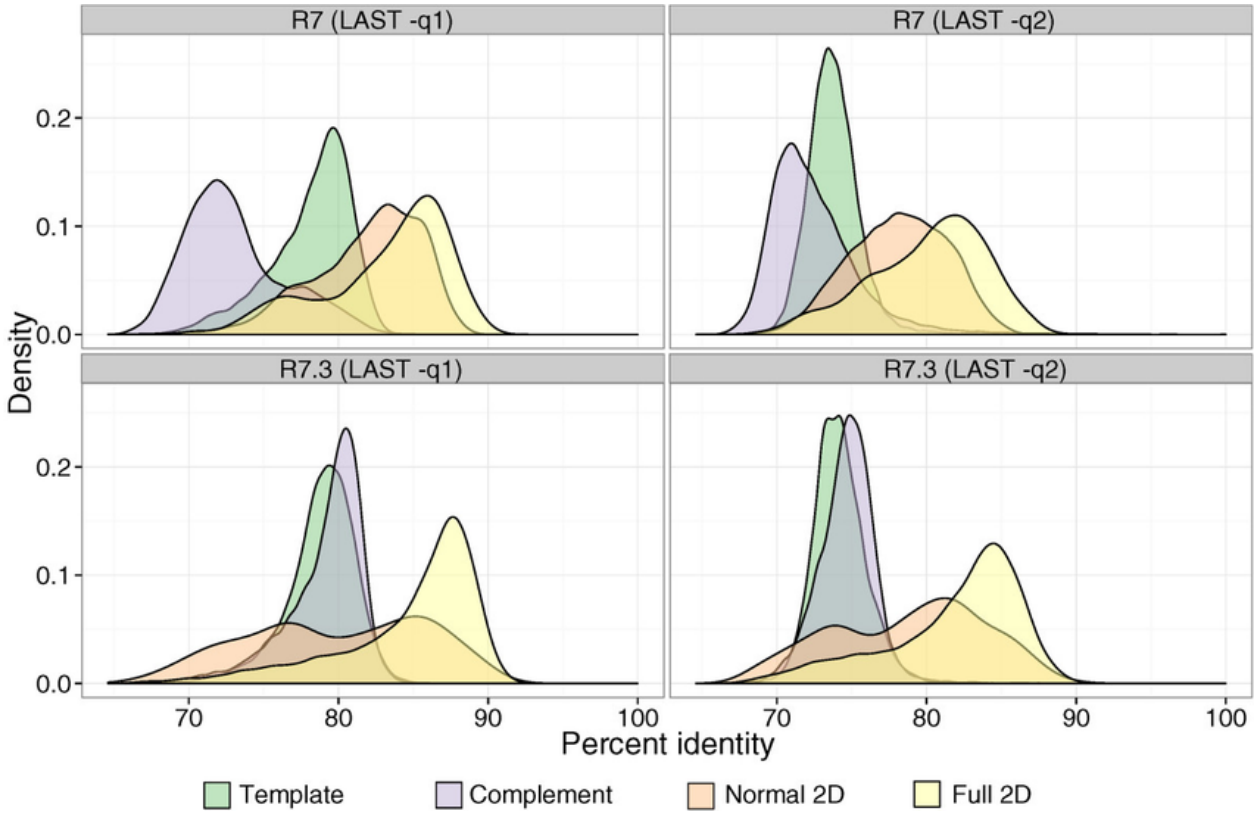
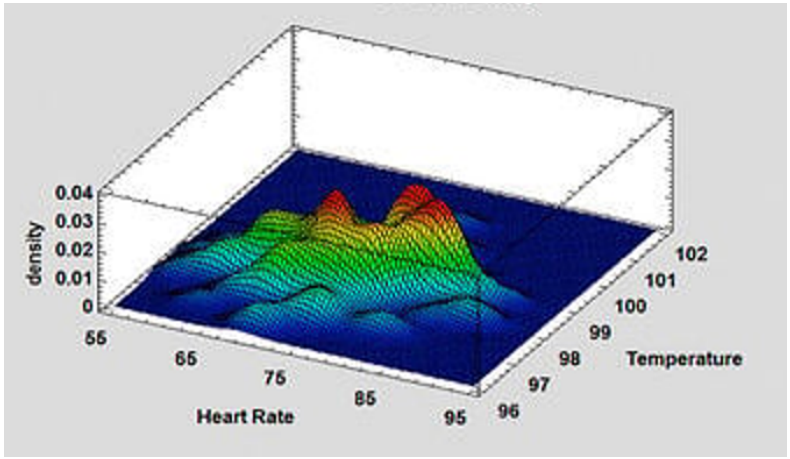
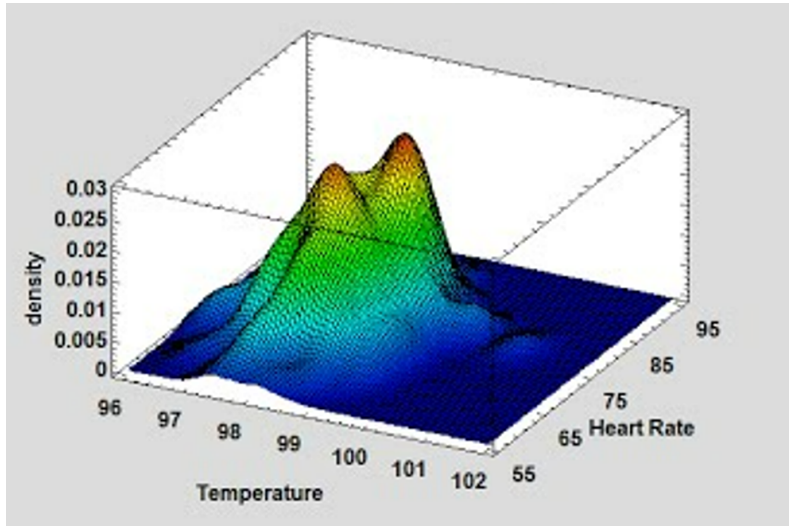
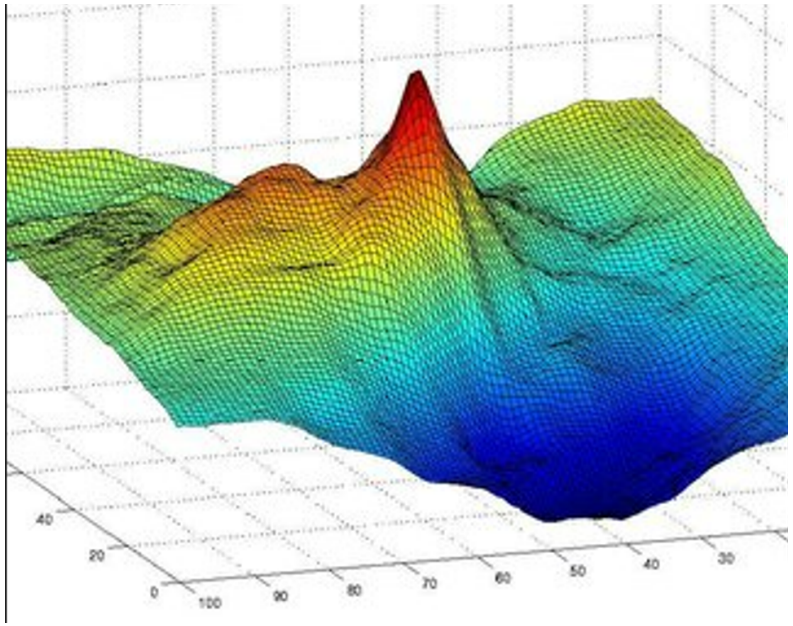


Figura: How to compare?

Comparando densidades multivariadas







Medidas de distância entre distribuições

- Existem muitas medidas de distância entre duas distribuições de probabilidade:
  - Kolmogorov: $\sup_x |F_1(x) - F_2(x)|$, onde F_i é a função de distribuição acumulada
  - Hellinger: $H(F_1, f_2) = \sqrt{0.5 \int (\sqrt{f_1(x)} - \sqrt{f_2(x)})^2 \, dx}$.
  - Variação total: $TV(f_1, f_2) = 0.5 \int |f_1(x) - f_2(x)| dx$ Pode-se mostrar quea distância da variação total pode ser escrita como $\sup_A \left| \int_A (f_1(x) - f_2(x)) dx \right| = \sup_A |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$: a maior diferença possível entre as probabilidades de A atribuídas por f_1 e f_2.
  - $\int |f_1(x) - f_2(x)| dx$
  - Existem várias conexões entre as diferentes distâncias. Por exemplo:

$$H^2(f_1, f_2) \leq TV(f_1, f_2) \leq \sqrt{2} H(F_1, f_2)$$

Medidas de distância entre distribuições

- Qualquer uma dessas medidas é intuitivamente razoável e poderia ser a base de uma teoria de seleção de modelos.
- Entretanto, a distância que gerou mais resultados teóricos e práticos foi a distância de Kullback-Leibler, definida a seguir E DENOTADA POR KL.

Suporte e surpresa

- Suponha que temos duas distribuições f_1 e f_2 competindo como modelos para descrever alguns dados.
- Seja $\mathcal{S}$ o suporte da distribuição f_1. Isto é,
  - No caso discreto: $\mathcal{S}_1 = \{z; \ \ \mathbb{P}_1(X = z) > 0\}$,
  - No caso contínuo, $\mathcal{S}_1 = \{z; \ \ f_1(z) > 0\}$
- Vamos assumir que $\mathcal{S}_1 = \mathcal{S}_2 = \mathcal{S}$.
- Para cada $z \in \mathcal{S}$, calculamos a surpresa ocasionada por gerar z sob f_1 e sob f_2.

Surpresa

- Para cada $z \in \mathcal{S}$, vemos a surpresa $-\log(f_i(z))$ de ocorrência sob f_1 e sob f_2.
- Se as duas distribuições são parecidas, esperamos que as surpresas $-\log(f_1(z))$ e $-\log(f_2(z))$ sejam similares para todo z.
- A diferença de surpresa é

$$-\log(f_2(z)) - (-\log(f_1(z))) = \log\left(\frac{f_1(z)}{f_2(z)}\right)$$

- Note que ela é o logaritmo da razão de verossimilhança do modelo 1 versus o modelo 2.
- Se $\log\left(f_1(z)/f_2(z)\right) = 0$, o valor z é igualmente provável nas duas distribuições
- Se $\log\left(f_1(z)/f_2(z)\right) > 0$, o valor z tem mais chance de ocorrer sob a distribuição 1.

Tirando a média

- Para ter uma idéia global ou um resumo da diferença de supresas considerando todos os possíveis valores z, tiramos uma "média" das diferenças $\log\left(f_1(z)/f_2(z)\right) > 0$.
- Queremos uma média ponderada sobre todos os valores possíveis de $z \in \mathcal{S}$.
- Queremos dar mais peso às discrepâncias $\log(f_X(z)/f_Y(z))$ associadas aos valores z que têm mais chance de ocorrer.
- Para isto, precisamos escolher um modelo de probabilidade para os valores $z \in \mathcal{S}$. Temos dois modelos possíveis: $f_1(z)$ ou $f_2(z)$.
- De maneira um tanto arbitrária, vamos escolher $f_1(z)$ para descrever as frequências dos valores $z \in \mathcal{S}$.

Kullback e Leibler (1951)

- A medida de Kullback-Leibler é definida no caso contínuo como

$$KL(f_1, f_2) = 2 \int_{\mathcal{S}} \log\left(\frac{f_1(z)}{f_2(z)}\right) f_1(z)\, dz = \mathbb{E}_{Z\sim f_1}\left(\frac{f_1(Z)}{f_2(Z)}\right)$$

- No caso discreto, como:

$$KL(p_1, p_2) = 2 \sum_{z_i \in \mathcal{S}} \log\left(\frac{\mathbb{P}_1(X = z_i)}{\mathbb{P}_2(X = z_i)}\right) \mathbb{P}_1(X = z_i) = \mathbb{E}_{Z\sim p_1}\left(\frac{p_1(Z)}{p_2(Z)}\right)$$

- Algumas vezes, a definição não utiliza a constante 2.

KL em passos

- Observe que KL é equivalente ao seguinte procedimento:
  - X é uma v.a. com densidade $f_1(x)$.
  - Ao gerar um valor aleatório X sob f_1, calcule a v.a. $Z = h(X) = 2\,\log(f_1(X)/f_2(X))$.
  - A seguir, tome esperança de Z (lembrando que X segue a distribuição f_1):

$$\begin{aligned} KL(f_1, f_2) &= \mathbb{E}_1[h(X)] \\ &= 2\,\mathbb{E}_1\left[\log\left(\frac{f_1(X)}{f_2(X)}\right)\right] \\ &= 2\int \log\left(\frac{f_1(x)}{f_2(x)}\right) \quad f_1(x)\,dx \end{aligned}$$

Exemplo: Poisson

- $X \sim$ Poisson(μ_1) e $Y \sim$ Poisson(μ_2).
- Temos

$$\log\left(\frac{p_X(k)}{p_Y(k)}\right) = \log\left(\frac{\mu_1^k \exp(-\mu_1)/k!}{\mu_2^k \exp(-\mu_2)/k!}\right) = k\ \log\left(\frac{\mu_1}{\mu_2}\right) - (\mu_1 - \mu_2)$$

- Por exemplo, se $\mu_1 = 3$ e $\mu_2 = 5$ então $\log(p_X(k)/p_Y(k)) = k\ \log(3/5) - (3-5) = -0.51\,k + 2$.
- Fazemos k aleatório com a distribuição de X e a medida de Kullback-Leibler é

$$\begin{aligned} KL(p_X, p_Y) &= 2\,E_X\left[X\ \log\left(\frac{\mu_1}{\mu_2}\right) - (\mu_1 - \mu_2)\right] \\ &= 2\left[E_X(X)\ \log\left(\frac{\mu_1}{\mu_2}\right) - (\mu_1 - \mu_2)\right] \\ &= 2\left[\mu_1\ \log\left(\frac{\mu_1}{\mu_2}\right) - (\mu_1 - \mu_2)\right] \end{aligned}$$

Exemplo: Poisson

- Vimos que, se $X \sim \text{Poisson}(\mu_1)$ e $Y \sim \text{Poisson}(\mu_2)$, então

$$KL(p_X, p_Y) = 2 \left[\mu_1 \log \left(\frac{\mu_1}{\mu_2} \right) - (\mu_1 - \mu_2) \right]$$

- Por exemplo, se $\mu_1 = 3$ e $\mu_2 = 5$, teremos $KL(p_X, p_Y) = 0.9350$.
- Se $\mu_1 = 30$ e $\mu_2 = 50$, teremos $KL(p_X, p_Y) = 9.3504$, o valor anterior multiplicado por 10.
- Se $\mu_1 = 0.3$ e $\mu_2 = 0.5$, teremos $KL(p_X, p_Y) = 0.093504$, o primeiro valor dividido por 10.
- De fato, é bem fácil mostrar que, se μ_1 e μ_2 são multiplicados por uma mesma constante $c > 0$, a distância KL entre duas Poissons também é multiplicada por c (basta olhar a fórmula).

Comparando Poissons

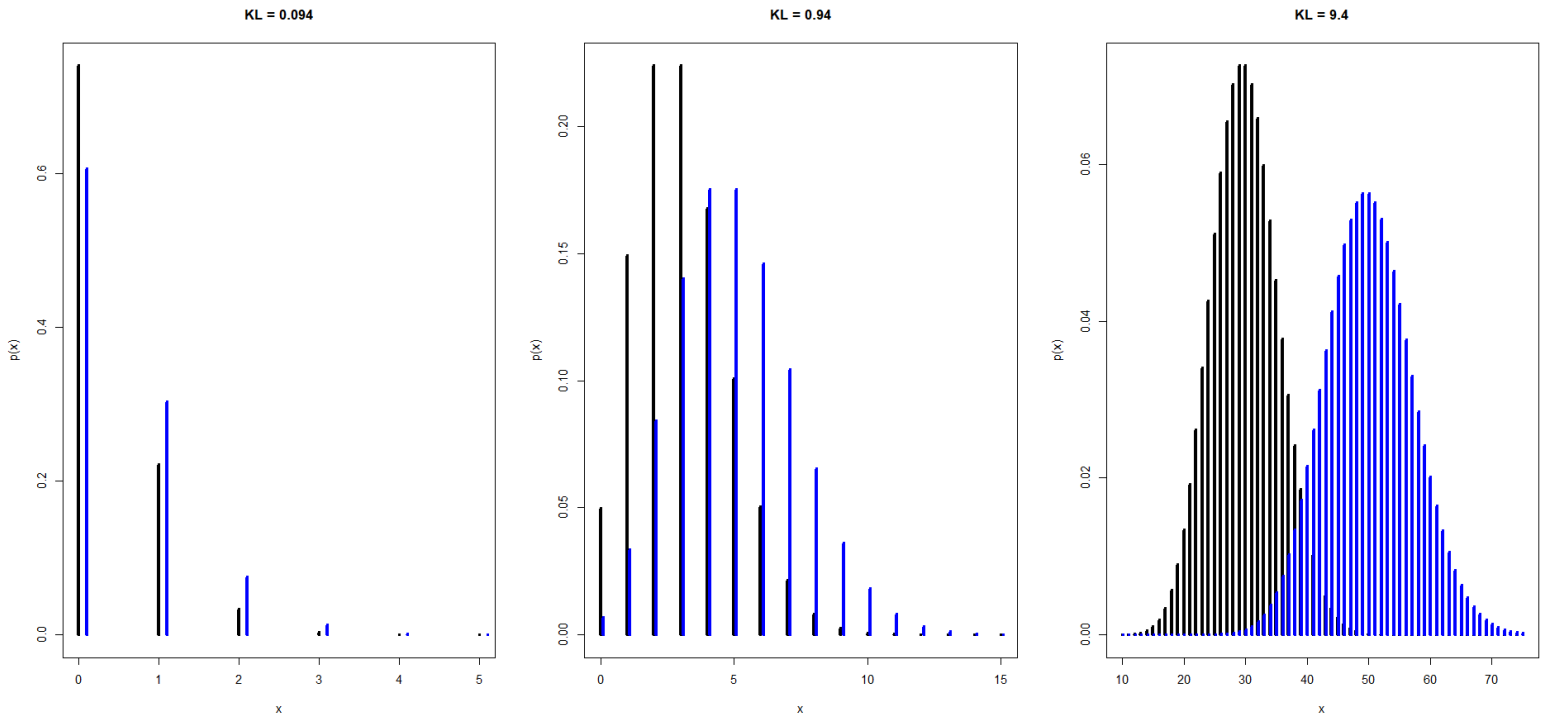


Figura: Different Poissons

KL não é rigorosamente uma distância

- Em geral, $KL(f_1, f_2) \neq KL(f_2, f_1)$.
- Isto é, KL não é simétrica nos seus argumentos.
- Isto implica que a medida de Kullback-Leibler não é, de fato, uma medida de distância no sentido matemático do termo.
- A distância KL de f_1 até f_2 não é igual à distância KL de f_2 até f_1

Poisson e distância *KL*

- Com $X \sim$ Poisson(μ_1) e $Y \sim$ Poisson(μ_2), temos

$$KL(p_X, p_Y) = 2 \left[\mu_x \log \left(\frac{\mu_x}{\mu_y} \right) - (\mu_x - \mu_y) \right]$$

- Se $\mu_x = 3$ e $\mu_y = 5$, teremos $KL(p_X, p_Y) = 0.9350$.
- Mas se trocarmos, fazendo $\mu_x = 5$ e $\mu_y = 3$, teremos $KL(p_X, p_Y) = 1.108256$.
- Vamos tentar entender o porquê dessa diferença daqui a pouco.

Distância *KL* é assimétrica

- No caso discreto, KL é definida assim:

$$KL(p_1, p_2) = \mathbb{E}_{Z \sim p_1} \log \left(\frac{p_1(Z)}{p_2(Z)} \right)$$

- Geramos um dado aleatório Z por p_1 (aqui está a assimetria)
- Em seguida, olhamos quão mais provável é este Z sob p_1 relativamente à p_2 calculando $p_1(Z)/p_2(Z)$
- Por exemplo, se $Z = z$ e $f_1(z)/f_2(z) = 3$, então o z gerado (por p_1) é 3 vezes mais provável sob p_1 do que sob p_2.
- Se $p_1 \approx p_2$, esperamos que essa razão $p_1(Z)/p_2(Z)$ seja tipicamente próxima de 1 para todo Z.
- Se p_1 for muito distante de p_2, esperamos que essa razão $p_1(Z)/p_2(Z)$ seja em geral bem maior que 1 (dado que Z veio de p_1).

Distância KL é assimétrica

- O fato é que olhamos a distância $KL(p_1, P_2)$ numa forma assimétrica.
- Um dado Z é gerado de p_1. Então $KL(p_1, p_2)$ mede o quanto podemos esperar que esse dado aleatório de p_1 pode ter sido gerado por p_2.
- $KL(p_2, p_1)$ mede a situação reversa: um dado Z é gerado por p_2 e nos pergutamos se é razoável que ele tenha vindo de p_1.
- O exemplo a seguir mostra que é razoável que $KL(p_2, p_1) \neq KL(p_1, p_2)$.

Distância *KL* é assimétrica

- Considere $f_1(x)$ uma uniforme $U(0,1)$ e $f_2(x)$ uma Beta$(20,20)$, bem concentrada em $(0.3, 0.7)$.

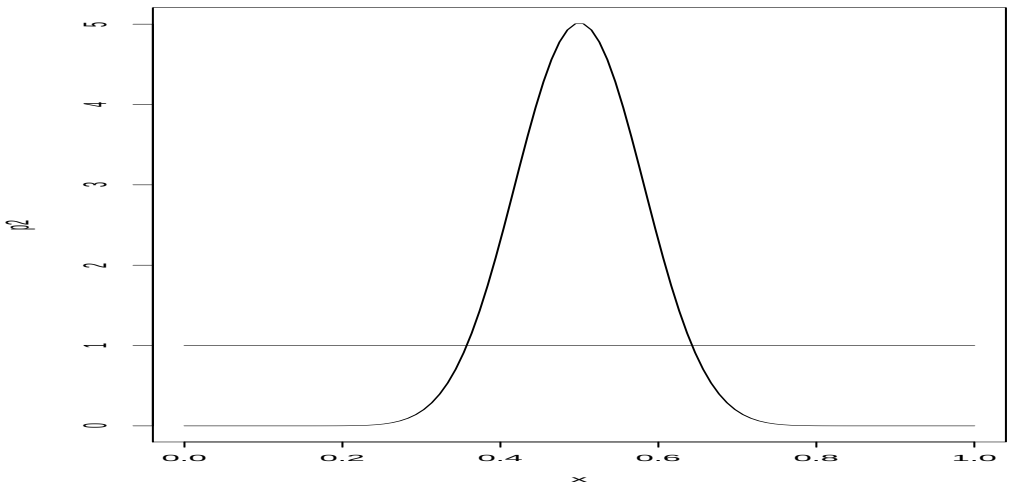


Figura: Densidades de $f_1(x) \sim U(0,1)$ e $f_2(x) \sim$ Beta$(20,20)$.

Distância *KL* é assimétrica

- $f_1(x) \sim U(0,1)$ e $f_2(x) \sim \text{Beta}(20,20)$, bem concentrada em $(0.3, 0.7)$.
- Um dado vindo de f_2 estará em geral bem concentrado em torno de $1/2$ e pode facilmente ser considerado como sendo gerado por uma $U(0,1)$. Por exemplo, se $Z = 0.4$ é gerado, poderíamos facilmente tomá-lo como tendo sido gerado por uma $U(0,1)$.
- Veja (a partir do gráfico) que $2\log(f_1(z)/f_2(z))$ fica entre $2\log(1/1) = 0$ e $2\log(5/1) = 3.2$.
- Podemos facilmente obter por simulação $KL(Beta(20,20), U(0,1)) \approx 2.2$.

Distância KL é assimétrica

- Vamos olhar agora a situação reversa: suponha que geramos $Z \sim U(0,1)$.
- Seria razoável esperar que este Z venha de uma Beta$(20,20)$?
- Se $Z \in (0.3, 0.7)$ n´ ao teremos razão para descrtar essa possibilidade.
- Entretanto, $\mathbb{P}(U(0,1) \notin (0.3,0.7)) = 0.6$.
- Isto é, existe uma chance razoável da $U(0,1)$ gerar um dado fora de $(0.3,0.7)$ onde a Beta$(20,20)$ está concentrada.
- Um dado fora de $(0.3,0.7)$ dificilmente seria gerado por uma Beta$(20,20)$.
- De fato, temos $KL(U(0,1), Beta(20,20)) \approx 20$ (por simulação)
- Isto é,
  $20 \approx KL(U(0,1), Beta(20,20)) >> KL(Beta(20,20), U(0,1)) \approx 2$

Caso normal

- Sejam $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}_1, \sigma_1^2 I_n)$ e $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}_2, \sigma_2^2 I_n)$.
- Neste caso,

$$\begin{aligned}\log\left(\frac{f_{\boldsymbol{X}}(\boldsymbol{z})}{f_{\boldsymbol{Y}}(\boldsymbol{z})}\right) &= \log\left(\frac{(2\pi\sigma_1^2)^{-n/2}\,\exp\left(\|\,\boldsymbol{z}-\boldsymbol{\mu}_1\|^2/(2\sigma_1^2)\right)}{(2\pi\sigma_2^2)^{-n/2}\,\exp\left(\|\,\boldsymbol{z}-\boldsymbol{\mu}_2\|^2/(2\sigma_2^2)\right)}\right) \\ &= n\log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2\sigma_1^2}\|\,\boldsymbol{z}-\boldsymbol{\mu}_1\|^2 + \frac{1}{2\sigma_2^2}\|\,\boldsymbol{z}-\boldsymbol{\mu}_2\|^2\end{aligned}$$

- Usando que $\|\,\boldsymbol{X}-\boldsymbol{\mu}_1\|^2/\sigma_1^2 \sim \chi^2(n)$ (qui-quadrado com n graus de liberdade) e que $\|\,\boldsymbol{X}-\boldsymbol{\mu}_2\|^2$ é uma qui-quadrado não-central podemos deduzir:

$$KL(f_{\boldsymbol{X}}, f_{\boldsymbol{Y}}) = 2\left[n\log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{n}{2} + \frac{n\sigma_1^2}{2\sigma_2^2} + \frac{\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2}{2\sigma_2^2}\right],$$

- É uma função da distância euclidiana entre os vetores de médias e das razões entre as variâncias das distribuições.

Caso normal

- Se $X \sim N(0,1)$, $Y \sim N(1,1)$ e $Z \sim N(5,1)$, então
- $KL(f_X, f_Y) = 1$, $KL(f_X, f_Z) = 25$ e $KL(f_Y, f_Z) = 16$
- Mas se agora tivermos $Z \sim N(0.5, 2^2)$, então
- $KL(f_X, f_Y) = 1$ (como antes) e $KL(f_X, f_Z) = 0.6988$ e $KL(f_Y, f_Z) = 0.6987$.
- Ou seja, Z está igualmente distante de X e Y e mais perto de cada uma delas que a distância entre X e Y.

Assimetria - Caso normal

- Seja $Y_1 \sim N(0, \sigma_1^2)$, $Y_2 \sim N(0, \sigma_2^2)$
- $$KL(f_1, f_2) = 2 \log \left(\frac{\sigma_2}{\sigma_1} \right) - 1 + \frac{\sigma_1^2}{\sigma_2^2}$$
- Considere $\sigma_1 = 1$ e $\sigma_2 = 5$
- Temos $KL(N(0,1), N(0,5^2)) = 2.26$
- mas $KL(N(0,5^2), N(0,1)) = 20.78$
- Todo dado vindo de uma $N(0,1)$ pode se passar como sendo gerado de uma $N(0,5^2)$
- Mas quase todo dado vindo de uma $KL(N(0,5^2)$ não consegue se passar como

KL divergência entre normais multivariadas gerais

- Sejam $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}_1, \Sigma_1)$ e $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}_2, \Sigma_2)$.
- A divergência de Kullback-Leibler entre $f_{\boldsymbol{X}}$ e $f_{\boldsymbol{Y}}$ é dada por:

$$KL(f_{\boldsymbol{X}}, f_{\boldsymbol{Y}}) = \frac{1}{2}\left[\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - n + \text{tr}(\Sigma_2^{-1}\Sigma_1) + \right.$$
$$\left. + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \Sigma_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right] .$$

- It is not symmetric

Jensen-Shannon

- Algumas vezes usamos uma outra medida (simétrica) chamada Jensen-Shannon divergence:

$$JS(f_1, f_2) = \frac{1}{2}\left(KL(f_1, f_m) + KL(f_2, f_m)\right)$$

onde

$$f_m(x) = (f_1(x) + f_2(x))/2\,.$$

é a mistura de f_1 e f_2.

- Continuaremos a usar $KL(f_1, f_2)$, chamando-a de distância entre as distribuições f_1 e f_2.

Local KL Divergence: A Quadratic Approximation

- Let $p(y;\boldsymbol{\theta})$ be a smooth parametric family of densities.
- We analyze the local behavior of the Kullback-Leibler divergence:

$$KL(p(y;\boldsymbol{\theta}) \,\|\, p(y;\boldsymbol{\theta}+\boldsymbol{\delta}))$$

  when δ is small.
- This KL measures the discrepancy between $p(y;\boldsymbol{\theta})$ and its perturbed version $p(y;\boldsymbol{\theta}+\boldsymbol{\delta})$.
- We will show that:

$$KL(p(y;\boldsymbol{\theta}) \,\|\, p(y;\boldsymbol{\theta}+\boldsymbol{\delta})) \approx \frac{1}{2}\boldsymbol{\delta}^\top \mathcal{I}(\boldsymbol{\theta})\boldsymbol{\delta}$$

  where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher Information Matrix.

Taylor Expansion of the Log-Likelihood

- The KL divergence can be rewritten as:

$$KL(p(y;\boldsymbol{\theta}) \,\|\, p(y;\boldsymbol{\theta}+\boldsymbol{\delta})) = \mathbb{E}_{\boldsymbol{\theta}}\left[\log \frac{p(Y;\boldsymbol{\theta})}{p(Y;\boldsymbol{\theta}+\boldsymbol{\delta})}\right]$$

- Since $\log p(Y;\boldsymbol{\theta}+\boldsymbol{\delta})$ is smooth, apply a second-order Taylor expansion:

$$\log p(Y;\boldsymbol{\theta}+\boldsymbol{\delta}) \approx \log p(Y;\boldsymbol{\theta}) + \boldsymbol{\delta}^\top \nabla_{\boldsymbol{\theta}} \log p(Y;\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\delta}^\top \nabla^2_{\boldsymbol{\theta}} \log p(Y;\boldsymbol{\theta})\boldsymbol{\delta}$$

- Substituting into the KL formula:

$$KL \approx \mathbb{E}_{\boldsymbol{\theta}}\left[-\boldsymbol{\delta}^\top \nabla_{\boldsymbol{\theta}} \log p(Y;\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\delta}^\top \nabla^2_{\boldsymbol{\theta}} \log p(Y;\boldsymbol{\theta})\boldsymbol{\delta}\right]$$

Expectations of the Expansion Terms

- The first-order term is the score and it has zero expectation:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\nabla_\theta \log p(Y;\boldsymbol{\theta})\right] = 0$$

- The expectation of the second-order term gives the Fisher Information:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[-\nabla^2_\theta \log p(Y;\boldsymbol{\theta})\right] = \mathcal{I}(\boldsymbol{\theta})$$

- Therefore:

$$KL(p(y;\boldsymbol{\theta}\|\boldsymbol{\theta}+\boldsymbol{\delta})) \approx \frac{1}{2}\boldsymbol{\delta}^\top \mathcal{I}(\boldsymbol{\theta})\boldsymbol{\delta}$$

- This is a second-order local approximation to the KL divergence.

Interpretation of the Approximation

- Note that the KL does NOT make use of the parametrization while the Fisher information does.
- The matrix $I(\theta)$ serves as a local metric on the parameter space.
- Small perturbations δ around θ result in quadratic increases in KL divergence.
- If $I(\theta)$ is large, then small changes in θ greatly affect the distribution.
- This result is central to:
  - Information geometry
  - Asymptotic theory in statistics
  - Bayesian inference (e.g., Laplace approximations)

KL Divergence Between Two Exponentials

- Let $f_1(y) = \lambda_1 e^{-\lambda_1 y}$ and $f_2(y) = \lambda_2 e^{-\lambda_2 y}$ be two exponential densities on $y > 0$.
- The Kullback-Leibler divergence from f_1 to f_2 is defined as:

$$KL(f_1 \| f_2) = \int_0^\infty f_1(y) \log\left(\frac{f_1(y)}{f_2(y)}\right) dy$$

- Plugging in the density functions:

$$KL(f_1 \| f_2) = \int_0^\infty \lambda_1 e^{-\lambda_1 y} \left[\log\left(\frac{\lambda_1}{\lambda_2}\right) + (\lambda_2 - \lambda_1) y\right] dy$$

KL Divergence Between Two Exponentials (cont.)

- Splitting the integral:

$$KL(f_1 \| f_2) = \log\left(\frac{\lambda_1}{\lambda_2}\right) \underbrace{\int_0^\infty \lambda_1 e^{-\lambda_1 y} dy}_{=1} + (\lambda_2 - \lambda_1) \underbrace{\int_0^\infty y \lambda_1 e^{-\lambda_1 y} dy}_{=\frac{1}{\lambda_1}}$$

- Final result:

$$KL(f_1 \| f_2) = \log\left(\frac{\lambda_1}{\lambda_2}\right) + \frac{\lambda_2}{\lambda_1} - 1$$

- This expression is always non-negative and equals 0 if and only if $\lambda_1 = \lambda_2$.

KL Divergence and information

- Let $f_1 = f_\lambda$ and $f_2 = f_{\lambda+\delta}$.

$$KL(f_\lambda \| f_{\lambda+\delta}) = \log\left(\frac{\lambda}{\lambda+\delta}\right) + \frac{\lambda+\delta}{\lambda} - 1$$

- We found that

$$KL(f_\lambda \| f_{\lambda+\delta}) \approx \frac{1}{2} \cdot \delta^2 \cdot \mathcal{I}(\lambda) = \frac{1}{2} \cdot \delta^2 \cdot \frac{1}{\lambda^2}$$

- Are they similar? Take $\lambda = 3$ and $\delta \in [-2, 2]$. We will plot $KL(f_\lambda \| f_{\lambda+\delta})$ and the approximation versus δ.

KL Divergence and information

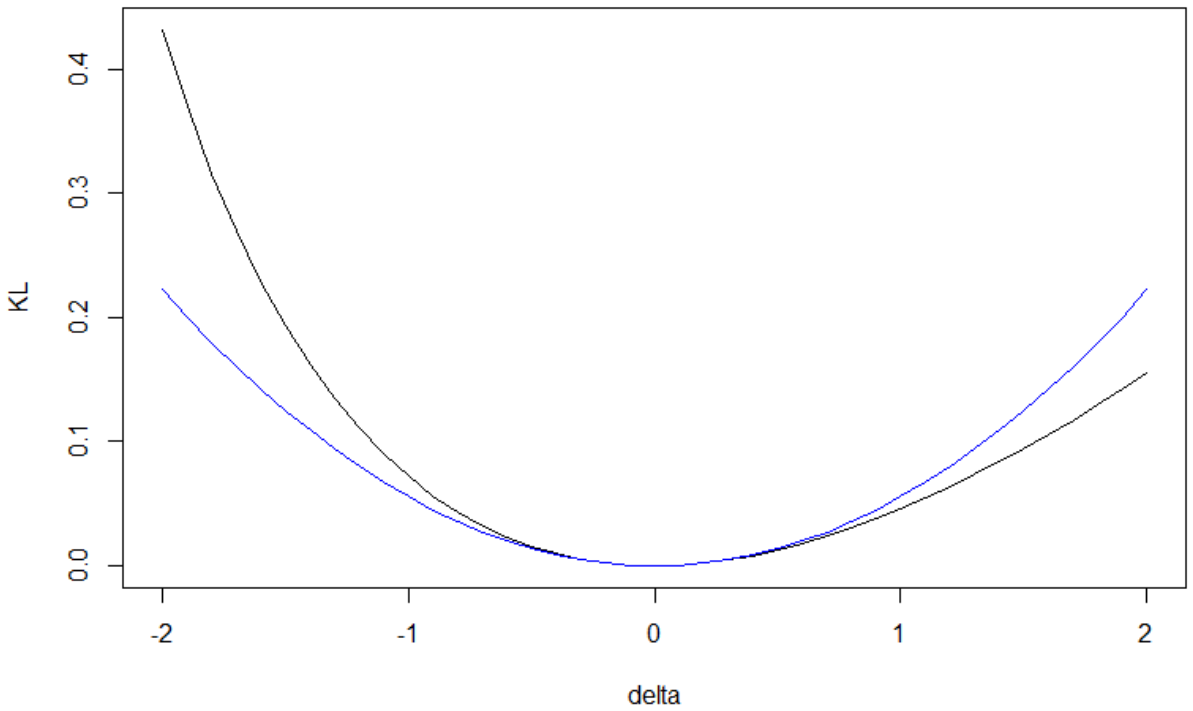


Figura: $KL(f_\lambda \| f_{\lambda+\delta}) \approx \frac{1}{2} \cdot \delta^2 \cdot \mathcal{I}(\lambda)$. Black: KL. Blue: Approximation with Information matrix

Relembrando de probab

- Se $Y_1, Y_2, \ldots, Y_n$ são i.i.d. com a mesma distribuição que Y e se $\mu = E(Y)$ então
$$\hat{\mu} = \sum_i Y_i/n \to \mu$$
- Isto é, o estimador $\sum_i Y_i/n$ (média aritmética dos elementos da amostra i.i.d.) converge para a a média populacional (a esperança) de Y.
- Isto vale PARA QUALQUER V.A. QUE POSSUA esperança.
- Repetindo: QUALQUER v.a. Y.
- Isto é muito simples mas muito importante quando acoplado com a idéia de transformar uma v.a. como veremos no próximo slide.

Relembrando de probab

- Se X é uma v.a. e $Y = h(X)$ é uma v.a. obtida como função de X.
- Por exemplo, $Y = X^2$ ou então $Y = \log(1 + X^2)$
- Se $X_1, \ldots, X_n$ são i.i.d. com distribuição de X então $Y_1 = h(X_1), \ldots, Y_n = h(X_n)$ são i.i.d. com uma distribuição de $Y = h(X)$.
- Por exemplo: $Y_1 = X_1^2, Y_2 = X_2^2, \ldots, Y_n = X_n^2$ são v.a.'s i.i.d.
- Outro exemplo: $Y_1 = \log(1 + X_1^2), Y_2 = \log(1 + X_2^2), \ldots, Y_n = \log(1 + X_n^2)$ são v.a.'s i.i.d.

Relembrando de probab

- Então $\mu_Y = E(Y) = E(h(X))$ pode ser estimado consistentemente pela média aritmática $\hat{\mu}_Y = \sum_i Y_i/n = \sum_i h(X_i)/n$.
- Por exemplo:

$$\frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \frac{X_1^2 + X_2^2 + \ldots + X_n^2}{n} \to \mathbb{E}(X^2) = \mathbb{E}(Y)$$

ou

$$\frac{Y_1 + \ldots Y_n}{n} = \frac{\log(1 + X_1^2) + \ldots + \log(1 + X_n^2)}{n} \to \mathbb{E}(\log(1 + X^2))$$

Estimando Kullback-Leibler

- A distância KL é apenas uma esperança de uma v.a. $Y = h(X)$ onde X é selecionada com a distribuição f_1:

$$KL(f_1, f_2) = 2\mathbb{E}_1 [h(X)] = 2\mathbb{E}_1 \left[\log \left(\frac{f_1(X)}{f_2(X)} \right) \right]$$

onde $h(x)$ é a função dada por

$$h(x) = \log \left(\frac{f_1(x)}{f_2(x)} \right)$$

- Se tivermos uma amostra i.i.d. de X obtida da distribuição f_1, podemos transformar cada X com a função h e a seeguir calcular a sua média aritmética.
- Este valor será aproximadamente igual a esperança teórica.

Estimando Kullback-Leibler

- Se $X_1, \ldots, X_n$ são i.i.d. com distribuição com desnidade f_1, podemos estimar

$$KL(f_1, f_2) = 2\mathbb{E}_1 \left[\log \left(\frac{f_1(X)}{f_2(X)} \right) \right]$$

pela média aritmética dos valores $h(X_i)$:

$$\widehat{KL(f_1, f_2)} = 2 \frac{1}{n} \sum_i \log \left(\frac{f_1(X_i)}{f_2(X_i)} \right)$$

- Nao podemos usar os dados vindos de f_1 para estimar desse modo simples a distancia $KL(f_2, f_1)$ pois precisamos de dados vindos de f_2 nesse caso.

Exemplo

- X e Y são v.a.'s Poisson com médias 3 e 5. Então

$$\log\left(\frac{p_X(k)}{p_Y(k)}\right) = k\log\left(\frac{3}{5}\right) - (3-5) = 2 - 0.511\,k$$

- Temos $KL(p_X, p_Y) = 2(2 - 0.511\,E_X(X)) = 2(0.468)$.
- Neste caso, sabemos exatamente qual é o valor de $KL(p_X, p_Y)$ e portanto, nem faz sentido estimá-lo. No entanto, se mesmo assim quiséssemos, podemos fazer o seguinte:
- Os valores observados de uma amostra de tamanho 10 de X com distribuição Poisson de média 3 são os seguintes:

$$5, 3, 4, 1, 2, 5, 2, 1, 8, 2$$

- Então $KL(p_X, p_Y) = 2(0.468)$ pode ser estimado por

$$\widehat{K(p_X, p_Y)} = \frac{2}{10}\sum_i (k_i \log(3/5) + 2) = 2(0.314)$$

Todo modelo é correto

- Em análise de dados com modelos paramétricos, selecionamos uma classe de distribuições $f(\boldsymbol{y}, \boldsymbol{\theta})$ para o vetor aleatório $\boldsymbol{Y}$.
- Fizemos uma análise de como o MLE $\widehat{\boldsymbol{\theta}}$ se comporta quando um dos elementos desta classe é o modelo gerador dos dados.
- Suponha que $\boldsymbol{\theta}_0 \in \Theta$ é o verdadeiro valor do parâmetro.
- Isto é, os dados são gerados pela distribuição $f(\boldsymbol{y}, \boldsymbol{\theta}_0)$
- Então, o MLE $\widehat{\boldsymbol{\theta}}$ baseado numa amostra possui as seguintes propriedades:
  - $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ quando $n \to \infty$ (é consistente)
  - $\mathbb{E}(\widehat{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}_0$ (é aproximadamente não-viciado)
  - $\mathbb{V}(\widehat{\boldsymbol{\theta}}) \approx \mathbb{I}^{-1}(\boldsymbol{\theta}_0)$
  - $\widehat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}_0, \mathbb{I}^{-1}(\boldsymbol{\theta}_0))$

Todo modelo é falso

- Uma frase famosa de George Box é: *Todo modelo é falso mas alguns são úteis*.
- Praticamente sempre, nossos modelos são distribuições idealizadas e simplificadoras. Esperamos que eles sejam capazes de descrever de forma aproximada o mecanismo verdadeiro que gera os dados.
- Mas se o modelo verdadeiro que gera os dados não é um membro $f(\boldsymbol{y}, \boldsymbol{\theta}_0)$ da classe $f(\boldsymbol{y}, \boldsymbol{\theta})$, então não existe nenhum $\boldsymbol{\theta}_0$ verdadeiro.
- Neste caso, o *MLE estará estimando o quê*?
- Ele converge para algum lugar quando a amostra aumenta?
- Se existir um ponto de convergência, ele faz algum sentido prático?
- A discussão a seguir vai supor que os dados sejam i.i.d. mas ela é válida num contexto mais geral de dados independentes mas não i.d. ou até mesmo de dados dependentes.

Minimizando *KL*

- Suponha que $g(y)$ é a verdadeira densidade que gera os dados observados.
- Propomos um modelo parametrizado como aproximação para g.
- O modelo é uma classe (um conjunto) de densidades $f(y, \boldsymbol{\theta})$ indexado pelo parâmetro $\boldsymbol{\theta}$.
- Vamos denotar por $f_{\boldsymbol{\theta}}$ a densidade $f(y, \boldsymbol{\theta})$.
- Uma estratégia interessante para modelar os dados gerados por g é procurar na classe de densidades $\{f(y, \boldsymbol{\theta})\}$ aquela que minimiza a distância *KL*.
- Isto é, vamos procurar na classe $\{f(y, \boldsymbol{\theta})\}$ aquele valor $\boldsymbol{\theta}_0$ tal que a densidade $f(y, \boldsymbol{\theta}_0)$ seja a mais próxima possível de g.
- Em símbolos,

$$\boldsymbol{\theta}_0 = \arg_{\boldsymbol{\theta}} \min KL(g, f_{\boldsymbol{\theta}})$$

Minimizando *KL*

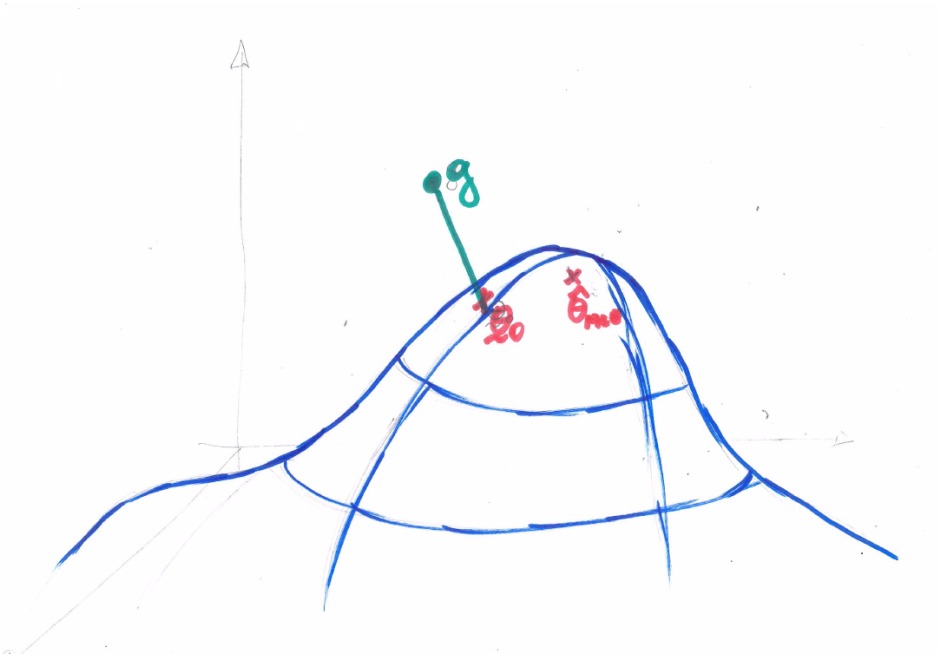


Figura: Cada ponto em $\mathbb{R}^3$ representa uma das infinitas distribuições de probabilidade existentes. A distribuição que gera os dados é representada pelo ponto g. A classe de distribuições do modelo $\{p(y, \theta)\}$ são os pontos na superfície curva, um conjunto de pontos bem restrito em $\mathbb{R}^3$. θ_0 é a distribuição mais próxima de g em termos da distância KL. També temos o MLE $\hat{\theta}$ como um ponto aleatório na superfície do modelo.

Minimizando *KL*

- Se $\boldsymbol{\theta}_0 = \arg_{\boldsymbol{\theta}} \min KL(g, f_{\boldsymbol{\theta}})$ então $f(y, \boldsymbol{\theta}_0)$ é o elemento da classe mais próximo de g.
- Se não pudermos encontrar g, o melhor que podemos fazer é usar $f(y, \boldsymbol{\theta}_0)$ em seu lugar.
- Temos

$$KL(g, f_{\boldsymbol{\theta}}) = \mathbb{E}_g \log\left(\frac{g(Y)}{f(Y, \boldsymbol{\theta})}\right) = \mathbb{E}_g \log\left(g(Y)\right) - \mathbb{E}_g \log\left(f(Y; \boldsymbol{\theta})\right)$$

- O primeiro termo, $\mathbb{E}_g(\log g(Y))$, não depende de $\boldsymbol{\theta}$.

Minimizando *KL*

- Repetindo:

$$KL(g, f_{\boldsymbol{\theta}}) = \mathbb{E}_g \log\left(\frac{g(Y)}{f(Y,\boldsymbol{\theta})}\right) = \mathbb{E}_g \log\left(g(Y)\right) - \mathbb{E}_g \log\left(f(Y;\boldsymbol{\theta})\right)$$

- Portanto, minimizar $KL(g, f_{\boldsymbol{\theta}})$ é o mesmo que procurar o valor de $\boldsymbol{\theta}$ que minimiza o segundo termo:

$$\boldsymbol{\theta}_0 = \arg_{\boldsymbol{\theta}} \min KL(g, f_{\boldsymbol{\theta}}) = \arg\min\left\{-\mathbb{E}_g \log\left(f(Y;\boldsymbol{\theta})\right)\right\}$$

- Esse segundo termo é a entropia cruzada de $f(Y;\boldsymbol{\theta})$ em relação a g.
- De forma equivalente, podemos maximizar o negativo desse segundo termos:

$$\boldsymbol{\theta}_0 = \arg\max \mathbb{E}_g \log\left(f(Y;\boldsymbol{\theta})\right)$$

Minimizando *KL*

- Como encontrar este elemento

$$\boldsymbol{\theta}_0 = \arg_{\boldsymbol{\theta}} \min KL(g, f_{\boldsymbol{\theta}}) = \arg\max \mathbb{E}_g \log\left(f(Y;\boldsymbol{\theta})\right) \quad ?$$

- Em alguns casos simples isto é possível, como veremos a seguir.
- g é uma densidade contínua arbitrária e que vai gerar os nossos dados.
- Nosso modelo é a classe de todas as gaussianas: $\{N(\mu, \sigma^2)\}$.
- Aqui $\boldsymbol{\theta} = (\mu, \sigma^2)$.
- Como encontrar (se é que existe) uma (única??) gaussiana $N(\mu_0, \sigma_0^2)$ que melhor aproxima uma densidade g arbitrária minimizando a distânca *KL*?
- Isto é, qual é a $N(\mu, \sigma^2)$ que tem $KL(g, N(\mu, \sigma^2))$ mínima?

Minimizando KL na classe das gaussianas

- Queremos (μ_0, σ_0^2) que minimizem

$$KL(g, N(\mu, \sigma^2)) = 2\mathbb{E}_g \log\left(g(X)\right) - 2\mathbb{E}_g \log\left(\phi(X; \mu, \sigma^2)\right)$$

onde $\phi(x; \mu, \sigma^2)$ é a densidade de uma gaussiana com parâmetros (μ, σ^2)

- O primeiro termo não depende de (μ, σ^2) e pode ser ignorado.
- Assim, basta maximizar

$$\begin{aligned} 2\mathbb{E}_g \log\left(\phi(X; \mu, \sigma^2)\right) &= 2\mathbb{E}_g \log\left((2\pi)^{-1/2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)\right. \\ &= \text{cte.} \; - \log\sigma^2 - \mathbb{E}_g\left(\frac{X-\mu}{\sigma}\right)^2 \\ &= \text{cte.} \; - \log\sigma^2 - \frac{1}{\sigma^2}\mathbb{E}_g\left(X-\mu\right)^2 \end{aligned}$$

Minimizando *KL* na classe das gaussianas

- Queremos (μ_0, σ_0^2) que maximizem

$$2\mathbb{E}_g \log\left(\phi(x;\mu,\sigma^2)\right) = \text{cte.} \;\; -\log\sigma^2 - \frac{1}{\sigma^2}\mathbb{E}_g\left(X-\mu\right)^2$$

- Para qualquer valor de σ^2 fixo, devemos minimizar $\mathbb{E}_g\left(X-\mu\right)^2$
- $\mathbb{E}_g\left(X-\mu\right)^2$ é minimizado se tomarmos $\mu_0 = \mathbb{E}_g(X)$.
- Com este valor μ_0 inserido na expressão acima, temos que achar σ^2 que maximize

$$\text{cte.} \;\; -\log\sigma^2 - \frac{1}{\sigma^2}\mathbb{E}_g\left(X-\mu_0\right)^2$$

- Derivando em σ^2 igualando a zero encontramos

$$\sigma_0^2 = \mathbb{E}_g\left(X-\mu_0\right)^2 = \mathbb{V}_g(X)$$

Minimizando *KL* na classe das gaussianas

- Isto é, dada uma g qualquer, a gaussiana $N(\mu, \sigma^2)$ que tem a distância de Kullback-Leibler mínima é $N(\mu_0, \sigma_0^2)$ onde $\mu_0 = \mathbb{E}_g(X)$ e $\sigma_0^2 = \mathbb{V}_g(X)$.
- Por exemplo, se g é a densidade de uma exponencial dupla (ou distribuição de Laplace, veja na wikipedia), com esperança 0 e variância 1 então a gaussiana que melhor aproxima é a $N(0, 1)$.

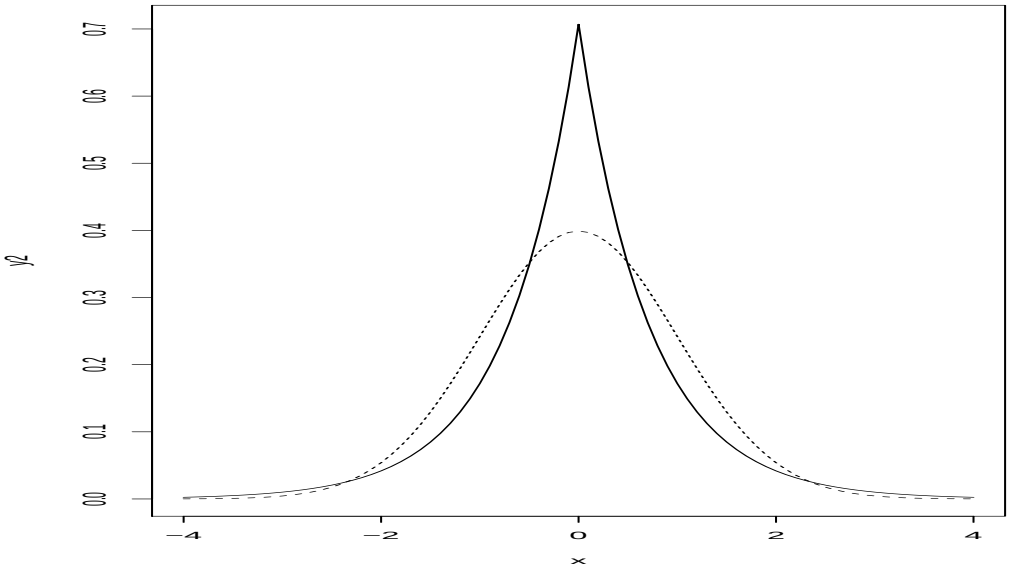


Figura: Laplace (linha contínua) e gaussiana mais próxima

Minimizando KL na classe das gaussianas

- Outro exemplo: g é a densidade de uma gama com $\alpha = 4$ e $\beta = 1$. Isto implica que g tem esperança 4 e variância 4.
- A gaussiana que melhor aproxima esta gama é a $N(4, 4)$.

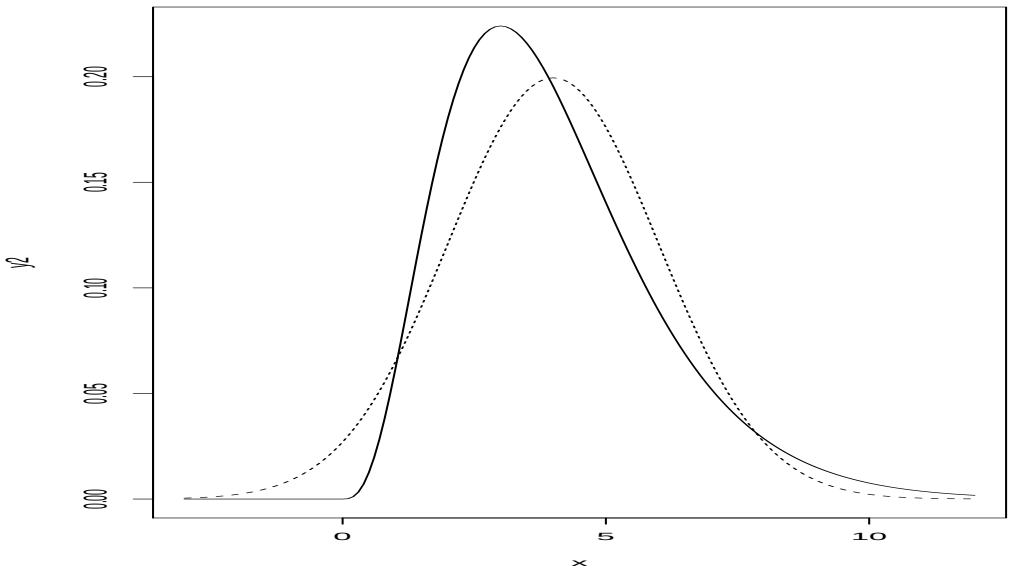


Figura: Gamma$(4, 1)$ (linha contínua) e gaussiana mais próxima $N(4, 4)$

Quando temos o modelo errado, o MLE estima o quê?

- Suponha que o vetor $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ é composto de v.a.'s i.i.d. com uma distribuição desconhecida com densidade $g(\boldsymbol{y})$.
- Adotamos um modelo $f(\boldsymbol{y}, \boldsymbol{\theta}) = \prod_i f(y_i, \boldsymbol{\theta})$ para os dados i.i.d. e obtemos o MLE maximizando a log-verossimilhança (dividida pela constante n):

$$\widehat{\boldsymbol{\theta}} = \arg\max \frac{1}{n} \sum_i \log f(Y_i; \boldsymbol{\theta})$$

O MLE estima o quê?

- Para cada valor $\boldsymbol{\theta}$ fixo, considere as v.a.'s $W_1 = \log f(Y_1; \boldsymbol{\theta}), \ldots, W_n = \log f(Y_n; \boldsymbol{\theta})$
- A média populacional (a esperança) de W é

$$\mathbb{E}_g(W) = \mathbb{E}_g\left(\log f(Y; \boldsymbol{\theta})\right)$$

  onde Y no lado direito é uma v.a. com densidade $g(y)$.
- Lembre-se de um resultado de probab (a lei dos grandes números): A média aritmética de v.a.'s i.i.d. converge para sua esperança populacional.
- A média aritmética baseada na amostra é

$$\frac{W_1 + \ldots + W_n}{n} = \frac{1}{n} \sum_i \log f(Y_i; \boldsymbol{\theta})$$

O MLE estima o quê?

- Pela Lei dos Grandes números, temos

$$\frac{W_1 + \ldots + W_n}{n} \to \mathbb{E}_g(W)$$

- Ou seja,

$$\frac{1}{n}\sum_i \log f(Y_i;\boldsymbol{\theta}) \to \mathbb{E}_g\left(\log f(Y;\boldsymbol{\theta})\right)$$

  para todo $\boldsymbol{\theta}$ fixo.
- Por definição o MLE de $\boldsymbol{\theta}$ é o argumento em $\boldsymbol{\theta}$ que maximiza a log-verossimilhança.
- Em símbolos:

$$\widehat{\boldsymbol{\theta}} = \arg_{\boldsymbol{\theta}} \max \frac{1}{n}\sum_i \log f(Y_i;\boldsymbol{\theta})$$

O MLE estima o quê?

- Vamos definir $\boldsymbol{\theta}_0$ como o argumento em $\boldsymbol{\theta}$ que maximiza

$$\boldsymbol{\theta}_0 = \arg_{\boldsymbol{\theta}} \max \mathbb{E}_g \left(\log f(Y; \boldsymbol{\theta}) \right)$$

- Mas nós acabamos de ver que maximizar $\mathbb{E}_g \left(\log f(Y; \boldsymbol{\theta}) \right)$ em $\boldsymbol{\theta}$ é o mesmo que minimizar $KL(g, f_{\boldsymbol{\theta}})$ em $\boldsymbol{\theta}$.
- Assim, $\boldsymbol{\theta}_0$ definido acima tem o memso significado que antes:

$$\boldsymbol{\theta}_0 = \arg_{\boldsymbol{\theta}} \max \mathbb{E}_g \left(\log f(Y; \boldsymbol{\theta}) \right) = \arg_{\boldsymbol{\theta}} \min KL(g, f(y; \boldsymbol{\theta}))$$

- Vamos agora relacionar o MLE $\widehat{\boldsymbol{\theta}}$ e $\boldsymbol{\theta}_0$.

O MLE estima o quê?

- O MLE $\widehat{\boldsymbol{\theta}}$ é uma v.a. e $\boldsymbol{\theta}_0$ é um valor fixo no espaço paramétrico, uma constante.
- Em geral, $\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_0$.
- Mas eles estão relacionados. Como:

$$\frac{1}{n}\sum_i \log f(Y_i;\boldsymbol{\theta}) \to \mathbb{E}_g\left(\log f(Y;\boldsymbol{\theta})\right)$$

podemos esperar que

$$\widehat{\boldsymbol{\theta}} = \arg_{\boldsymbol{\theta}} \max \frac{1}{n}\sum_i \log f(Y_i;\boldsymbol{\theta}) \to \arg_{\boldsymbol{\theta}} \max \mathbb{E}_g\left(\log f(Y;\boldsymbol{\theta})\right) = \boldsymbol{\theta}_0$$

- De fato, podemos demonstrar isto rigorosamente sob certas condições mas não faremos isto neste curso.

O MLE estima o quê?

- Assim, temos

$$\widehat{\boldsymbol{\theta}} = \arg_{\boldsymbol{\theta}} \max \frac{1}{n} \sum_i \log f(Y_i; \boldsymbol{\theta}) \to \arg_{\boldsymbol{\theta}} \max \mathbb{E}_g \left(\log f(Y; \boldsymbol{\theta}) \right) = \boldsymbol{\theta}_0$$

- A medida que n cresce, o MLE $\widehat{\boldsymbol{\theta}}$ converge para o valor $\boldsymbol{\theta}_0$ que minimiza a distância KL entre o modelo verdadeiro g e a classe $\{f(y, \boldsymbol{\theta})\}$.
- Agora entedemos o que o MLE está fazendo.
- Existe um elemento $\boldsymbol{\theta}_0$ da classe de distribuições $\{f(y, \boldsymbol{\theta})\}$ que forma nosso modelo que é a mais KL-próxima possível da distribuição verdadeira g.
- O MLE $\widehat{\boldsymbol{\theta}}$ é um estimador deste valor $\boldsymbol{\theta}_0$.

O MLE estima o quê?

- Se g for de fato um elemento $f(y, \boldsymbol{\theta}_0)$ da classe especificada no modelo, temos todos os resultados que já vimos neste curso:

$$\widehat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}_0, I^{-1}(\boldsymbol{\theta}_0)$$

- No caso mais comum em que g não pertence à classe $\{f(y, \boldsymbol{\theta})\}$ do modelo, Peter Huber (1967) demonstrou que, se a amostra não é muito pequena:

$$\widehat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}_0, V)$$

onde V mistura a informação de Fisher com outra matriz.

O MLE estima o quê?

- MOre specifically:

$$\widehat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}_0, V)$$

onde V é chamada de variancia sanduiche e é igual a

$$V = H^{-1} J H^{-1}$$

com

$$H = \mathbb{E}_g \left[\nabla^2_{\boldsymbol{\theta}} \log f(Y; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right] \quad \text{(Hessiana esperada)}$$

$$J = \mathbb{E}_g \left[\nabla_{\boldsymbol{\theta}} \log f(Y; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log f(Y; \boldsymbol{\theta})^\top \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right] \quad \text{(score “quadrado”)}$$

- Quando o modelo é correto, teremos $H = J = \mathcal{I}(\boldsymbol{\theta})$ e V será a informação de Fisher usual: $V = \mathcal{I}(\boldsymbol{\theta})$.

Estimando H e J a partir dos dados

- Embora as esperanças definindo H e J sejam em relação à distribuição verdadeira g, podemos estimá-las usando a distribuição empírica dos dados e o modelo adotado $f(y;\boldsymbol{\theta})$.
- As derivadas são tomadas em relação a $\boldsymbol{\theta}$, mas avaliadas no ponto $\boldsymbol{\theta}_0$ (desconhecido), que estimamos por $\widehat{\boldsymbol{\theta}}$.
- Estimativas:

$$\widehat{J} = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log f(Y_i;\widehat{\boldsymbol{\theta}}) \cdot \nabla_{\boldsymbol{\theta}} \log f(Y_i;\widehat{\boldsymbol{\theta}})^\top$$

$$\widehat{H} = \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \log f(Y_i;\widehat{\boldsymbol{\theta}})$$

- Ambas são obtidas numericamente a partir da função de log-verossimilhança.

O estimador sanduíche da variância

- Com as estimativas anteriores, construímos a matriz de variância robusta (ou "sanduíche"):

$$\widehat{V} = \widehat{H}^{-1}\,\widehat{J}\,\widehat{H}^{-1}$$

- Essa matriz estima a variância assintótica do MLE mesmo quando o modelo está incorretamente especificado.
- É muito usada em:
  - Modelos de verossimilhança mal especificados
  - Equações de estimação generalizadas (GEEs)
  - Econometria (ex: correção de heterocedasticidade de White)
- O mais incrível neste processo é que podemos obter o MLE e uma estimativa $f(y, \widehat{\boldsymbol{\theta}})$ da melhor aproximação $f(y, \boldsymbol{\theta}_0)$ de g sem ter a menor idéia de quem é g.
- É uma ferramenta poderosa: permite inferência confiável mesmo quando não conhecemos a distribuição verdadeira g.

Comparando modelos

- Suponha que temos dois modelos alternativos, 1 e 2.
- Cada um deles é uma classe de distribuições indexadas por parâmetros.
- Digamos, θ para o modelo 1 e ϕ para o modelo 2.
- Modelo 1: $f(y, \theta)$.
- Modelo 2: $h(y; \phi)$ (vamos usar h para as densidades do modelo 2)
- Os parâmetros θ e ϕ podem ter dimensões e interpretações físicas diferentes.
- Qual deles é o melhor para descrever dados gerados de uma distribuição g desconhecida?

Comparando modelos

- Temos uma medida de distância entre g e uma classe de distribuições.
- Seja θ_0 o valor de θ que minimiza a distância $KL(g, f(y, \theta))$.
- Isto é,

$$\theta_0 = \arg\min KL(g, f(y, \theta))$$

- A distância mínima entre g e o modelo 1 é $KL(g, f(y, \theta_0))$.
- Do mesmo modo, teremos a distância mínima entre g e o modelo 2 dada por $KL(g, h(y, \phi_0))$ onde

$$\phi_0 = \arg\min KL(g, h(y, \phi))$$

Comparando modelos

- Assim, o natural é comparar $KL(g, f(y, \theta_0))$ e $KL(g, h(y, \phi_0))$.
- O que tiver distância mínima é o escolhido.
- Podemos acrescentar um critério adicional: Se o modelo 2 for muito mais complicado que o modelo 1 e se $KL(g, f(y, \theta_0)) > KL(g, h(y, \phi_0))$ mas a diferença for muito pequena, podemos ficar com o modelo 1, mais simples e que tem praticamente a mesma distância que o modelo 2.
- OK, mas como definir se a distância é pequena?
- E muito mais importante: como calcular $KL(g, f(y, \theta_0))$ e $KL(g, h(y, \phi_0))$ na prática?

Akaike

- Akaike resolveu estes dois problemas para nós.
- A segunda pergunta é fácil: como calcular $KL(g, f(y, \theta_0))$ e $KL(g, h(y, \phi_0))$ na prática?
- Como o MLE $\hat{\boldsymbol{\theta}} \to \theta_0$ e o MLE $\hat{\boldsymbol{\phi}} \to \phi_0$, o natural é substituir θ_0 e ϕ_0 pelos seus MLEs e comparar $KL(g, f(y, \hat{\boldsymbol{\theta}}))$ e $KL(g, h(y, \hat{\boldsymbol{\phi}}))$.
- Como vimos antes, essas estimativas são simplesmente as log-verossimilhanças de cada modelo no seu valor máximo.
- Isto é, nesta abordagem, bastaria comparar as verosmilhanças maximizadas de cada modelo.
- Entretanto, Akaike mostrou que $KL(g, f(y, \hat{\boldsymbol{\theta}}))$ não é uma boa estimativa de $KL(g, f(y, \theta_0))$.

Akaike

- Sabemos que $KL(g, f(y, \hat{\theta})) > KL(g, f(y, \hat{\theta}))$ pois θ_0 é o minimizador do KL.
- Existe um vício positivo: $\hat{\theta}$ é aleatório e mostra-se que o valor esperado de $KL(g, f(y, \hat{\theta}))$ é maior que $KL(g, f(y, \hat{\theta}))$:

$$\mathbb{E}_g \left[KL(g, f(y, \hat{\theta})) \right] > KL(g, f(y, \hat{\theta}))$$

- Este vício é causado por over-fitting: estamos usando os dados de treino para estimar o modelo e tambéd para avaliarmos qual modelo é melhor.
- Akaike encontrou uma fórmula para este vício de over-fitting e com isso corrigiu $KL(g, f(y, \hat{\theta}))$ e $KL(g, h(y, \hat{\phi}))$.

Akaike

- Seja k o índice do modelo ($k = 1$ ou $k = 2$, em nosso exemplo, mas podemos ter vários modelos ao mesmo tempo)
- Seja p_k o número de parâmetros livres do modelo k.
- Ele mostrou que, se calcularmos o valor de

$$AIC(k) = -2 \log f(\boldsymbol{y}, \widehat{\theta}_k) + 2p_k$$

para cada modelo alternativo, o que tiver o menor $AIC(k)$ deve ser o melhor modelo.

Usando o AIC na prática

- Quando temos vários modelos candidatos, com diferentes estruturas ou números de variáveis explicativas, o AIC fornece uma regra objetiva para comparação.
- Para cada modelo k, calculamos:

$$AIC(k) = -2 \log f(\boldsymbol{y}; \widehat{\theta}_k) + 2p_k$$

- O modelo com menor valor de AIC é considerado o melhor.
- Podemos comparar modelos com diferentes conjuntos de variáveis, diferentes distribuições, ou até diferentes formas funcionais.
- Importante: todos os AICs devem ser calculados com a mesma resposta $\boldsymbol{y}$ sobre o mesmo conjunto de dados.

Exemplo: regressão linear múltipla

- Suponha que queremos modelar o consumo de energia Y com base em variáveis como temperatura (X_1), umidade (X_2), e velocidade do vento (X_3).
- Ajustamos três modelos lineares:
  - Modelo 1: $Y = \beta_0 + \beta_1 X_1$
  - Modelo 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - Modelo 3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- A log-verossimilhança de cada modelo avaliada no MLE será proporcional ao MSE de cada modelo.
- Se olharmos o MSE dos três modelos o modelo mais completo terá menor MSE.
- Por quê?

Exemplo: regressão linear múltipla

- Ajustamos três modelos lineares:
  - Modelo 1: $Y = \beta_0 + \beta_1 X_1$
  - Modelo 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - Modelo 3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- Calculamos o AIC para os três modelos:

$$AIC(k) = -2 \log f(\boldsymbol{y}; \widehat{\theta}_k) + 2p_k$$

- Mesmo que o Modelo 3 tenha maior log-verossimilhança, o AIC pode indicar que o Modelo 2 é melhor (trade-off ajuste vs. complexidade).
- Isso evita overfitting e melhora a capacidade preditiva em novos dados.

Exemplo: regressão logística

- Agora, a variável resposta Y indica se houve ou não falha em um equipamento: $Y = 1$ (falha), $Y = 0$ (sem falha).
- Ajustamos três modelos logísticos:
  - Modelo 1: $\log \frac{p}{1-p} = \beta_0 + \beta_1 X_1$
  - Modelo 2: $\log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - Modelo 3: $\log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- Para cada modelo, usamos a log-verossimilhança da regressão logística e computamos o AIC.
- Mesmo raciocínio: menor AIC $\rightarrow$ melhor equilíbrio entre ajuste e complexidade.
- A abordagem funciona mesmo com modelos de natureza não linear.

Ideia teórica por trás do AIC

- O objetivo de Akaike era comparar os valores mínimos de $KL(g, f(y;\theta))$ para diferentes modelos.
- Como vimos:

$$KL(g, f(y;\theta)) = \mathbb{E}_g[\log g(Y)] - \mathbb{E}_g[\log f(Y;\theta)]$$

- Como $\mathbb{E}_g[\log g(Y)]$ é constante (não depende do modelo), basta comparar:

$$-\min_\theta \mathbb{E}_g[\log f(Y;\theta)] = -\mathbb{E}_g[\log f(Y;\theta_0)]$$

de cada modelo.

- Mas este valor não é observável diretamente pois não conhecemos θ_0.
- temos apenas uma estimativa de θ_0 baseada no MLE $\hat{\theta}$.

Como Akaike resolveu os dois problemas?

- Akaike mostrou que o valor observado da log-verossimilhança avaliada no MLE:

$$\log f(\boldsymbol{y}; \widehat{\theta})$$

é uma estimativa viciada de

$$\mathbb{E}_g[\log f(Y; \theta_0)]$$

onde θ_0 é o valor ótimo de acordo com KL.

- Defina o viés:

$$\mathbb{E}_g[\log f(\boldsymbol{y}; \widehat{\theta})] - \mathbb{E}_g[\log f(Y; \theta_0)]$$

- Ele obteve uma correção assintótica para esse viés, proporcional ao número de parâmetros p_k do modelo.

Como Akaike corrigiu o viés?

- Akaike mostrou que, assintoticamente, esse viés é aproximadamente igual a:

$$\text{Viés} \approx p_k$$

onde p_k é o número de parâmetros do modelo.

- Isso significa que a log-verossimilhança observada tende a superestimar o desempenho preditivo fora da amostra.
- Para corrigir esse viés, Akaike propôs o **Akaike Information Criterion (AIC)**:

$$\boxed{\text{AIC} = -2 \log f(\boldsymbol{y}; \widehat{\theta}) + 2p_k}$$

- O termo $2p_k$ penaliza o excesso de complexidade e ajusta a superestimação da log-verossimilhança.
- O AIC permite comparar modelos com diferentes números de parâmetros, favorecendo os mais parcimoniosos com bom ajuste.

Como Akaike resolveu os dois problemas?

- Assim nasceu o critério:

$$AIC(k) = -2 \log f(\boldsymbol{y}; \widehat{\theta}_k) + 2p_k$$

- O AIC é uma estimativa assintoticamente não-viesada de $-2 \cdot \mathbb{E}_g[\log f(Y; \theta_0)]$, onde $\widehat{\theta}$ é uma aproximação de θ_0 obtida via MLE.
- O AIC aproxima $-2 \cdot \mathbb{E}_g[\log f(Y; \theta_0)]$, isto é, mede a perda de informação esperada ao usar o modelo $f(y; \theta)$ no lugar de g.
- O modelo com menor AIC é o mais próximo de g em termos de KL.

Otimismo da log-verossimilhança

- Mesmo quando o modelo está corretamente especificado (isto é, $g = f(y;\theta_0)$ para algum θ_0), o MLE maximiza a verossimilhança:

$$L(\widehat{\theta}) = \log f(\boldsymbol{y};\widehat{\theta}) \geq \log f(\boldsymbol{y};\theta_0)$$

- Isso acontece porque $\widehat{\theta}$ foi escolhido para maximizar a verossimilhança nos dados observados.
- Assim, o valor observado $L(\widehat{\theta})$ é um estimador viciado para cima de $\mathbb{E}_g[\log f(Y;\theta_0)]$.
- Akaike mostrou que essa mesma ideia vale mesmo quando o modelo está mal especificado, e a distribuição verdadeira g não pertence à família $\{f(y;\theta)\}$.

Esboço da demonstração de Akaike

- $Y_1, \ldots, Y_n$ são i.i.d. com distribuição verdadeira g.
- Defina o MLE:

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(Y_i; \theta)$$

- Nosso objetivo é estimar:

$$-2 \cdot \mathbb{E}_g[\log f(Y; \theta_0)]$$

- Mas só temos acesso a $L(\widehat{\theta}) = \log f(\boldsymbol{y}; \widehat{\theta})$.
- Usando expansão de Taylor de segunda ordem de $\log f(Y; \widehat{\theta})$ em torno de θ_0, e resultados assintóticos do MLE, vamos mostrar que:

$$\mathbb{E}_g\left[\log f(\boldsymbol{y}; \widehat{\theta})\right] = n \cdot \mathbb{E}_g[\log f(Y; \theta_0)] + p + o(1)$$

Expansão de Taylor da log-verossimilhança

- Vamos expandir $\log f(Y;\widehat{\theta})$ em torno de θ_0 usando Taylor de segunda ordem:

$$\log f(Y;\widehat{\theta}) \approx \log f(Y;\theta_0) + (\widehat{\theta}-\theta_0)^\top \nabla_\theta \log f(Y;\theta_0) + \frac{1}{2}(\widehat{\theta}-\theta_0)^\top \nabla^2_\theta \log f(Y;\theta_0)(\widehat{\theta}-\theta_0)$$

- vamos poder ignorrar o segundo termo, como explicamso agora.
- A expectativa de $\nabla_\theta \log f(Y;\theta_0)$ sob g é zero (por definição de θ_0).
- Vamos demosntrar isso a seguir.

Mesmo com modelo errado: o score tem média zero

- Seja $g(y)$ a densidade verdadeira e $f(y;\theta)$ a densidade do modelo.
- Defina:

$$\ell(\theta) = \mathbb{E}_g[\log f(Y;\theta)] = \int \log f(y;\theta) \cdot g(y)\, dy$$

- Suponha que $\ell(\theta)$ é diferenciável e que podemos trocar derivada e integral:

$$\nabla_\theta \ell(\theta) = \int \nabla_\theta \log f(y;\theta) \cdot g(y)\, dy$$

- Seja $\theta_0 = \arg\max_\theta \ell(\theta)$, ou seja, o valor que minimiza a divergência KL entre g e o modelo.
- Então:

$$\mathbb{E}_g[\nabla_\theta \log f(Y;\theta_0)] = \nabla_\theta \ell(\theta_0) = 0$$

- Mesmo com o modelo errado, o score tem esperança nula em θ_0.

Por que o termo linear da expansão pode ser ignorado?

- Na expansão de Taylor de $\log f(Y;\widehat{\theta})$ em torno de θ_0 aparece o termo:

$$(\widehat{\theta} - \theta_0)^\top \nabla_\theta \log f(Y;\theta_0)$$

- Este é o produto de dois termos aleatórios, e não podemos separar a esperança:

$$\mathbb{E}_g[(\widehat{\theta} - \theta_0)^\top \nabla_\theta \log f(Y;\theta_0)] \neq (\mathbb{E}_g[\widehat{\theta} - \theta_0])^\top \cdot \mathbb{E}_g[\nabla_\theta \log f(Y;\theta_0)]$$

- Porém:
  - $\widehat{\theta} - \theta_0 = \mathcal{O}_p(n^{-1/2})$
  - $\nabla_\theta \log f(Y;\theta_0)$ é $\mathcal{O}_p(1)$ e tem média zero
  - Os dois termos são assintoticamente (quase) independentes
- Resultado:

$$\mathbb{E}_g\left[(\widehat{\theta} - \theta_0)^\top \nabla_\theta \log f(Y;\theta_0)\right] = o(n^{-1})$$

- Podemos ignorar esse termo na análise assintótica do viés do AIC.

Consequência da expansão de Taylor

- A variância de $\widehat{\theta}$ em torno de θ_0 é da ordem de $1/n$ e tende para a inversa da informação de Fisher.
- Assim, ao tomar a esperança em g, o primeiro termo de ordem não nula vem do termo quadrático da expansão.
- Para o segundo termo, temos uma forma quadratica que se aproxima de uma distribuicao qui-quadrado. Seu valor esperado e' a dimensao do vetor.

Consequência da expansão de Taylor

- Assim, a expectativa da log-verossimilhança avaliada em $\widehat{\theta}$ é então:

$$\mathbb{E}_g[\log f(Y;\widehat{\theta})] \approx \mathbb{E}_g[\log f(Y;\theta_0)] + \frac{p}{n}$$

- Multiplicando por n, temos:

$$\mathbb{E}_g[\log f(\boldsymbol{y};\widehat{\theta})] \approx n \cdot \mathbb{E}_g[\log f(Y;\theta_0)] + p$$

- Logo, o valor observado da log-verossimilhança é otimista, ele está em média p unidades acima do valor esperado verdadeiro.
- Esse viés é justamente o que é compensado pelo termo $+2p$ no AIC.

Conclusão: o viés e o AIC

- O viés é, aproximadamente:

$$\mathbb{E}_g[\log f(\boldsymbol{y};\widehat{\theta})] - n \cdot \mathbb{E}_g[\log f(Y;\theta_0)] \approx p$$

- Multiplicando por -2, temos:

$$-2 \cdot \log f(\boldsymbol{y};\widehat{\theta}) + 2p \approx -2 \cdot \mathbb{E}_g[\log f(Y;\theta_0)] + \text{constante}$$

- Portanto, o critério

$$AIC = -2 \cdot \log f(\boldsymbol{y};\widehat{\theta}) + 2p$$

é uma estimativa assintoticamente não-viesada (até constante aditiva) de $-2 \cdot \mathbb{E}_g[\log f(Y;\theta_0)]$.

- Mínimo AIC $\rightarrow$ modelo mais próximo da distribuição verdadeira g em termos de KL.