

AGA 0505 - Análise de Dados em Astronomia

6. O Método da Máxima Verossimilhança

Laerte Sodré Jr.

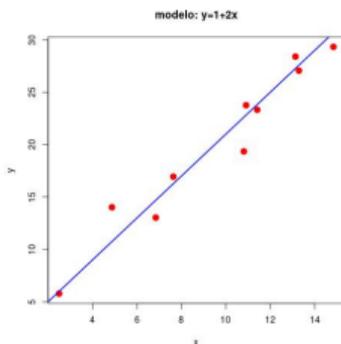
1o. semestre, 2024

aula de hoje:

1. modelagem dos dados
 2. a verossimilhança
 3. inferência de parâmetros pelo método da máxima verossimilhança
 4. exemplo: combinação de medidas
 5. exemplo: distribuição binomial
 6. exemplo: distribuição poissoniana
 7. exemplo: ajuste de funções com erros gaussianos
 8. otimização de funções
 9. incertezas nos parâmetros
 10. regressão linear
 11. regressão multi-linear
 12. modelos não-lineares: perfil de brilho de NGC 4472 (M49)
- Essencialmente todos os modelos estão errados, mas alguns são úteis*
- George Box*

modelagem dos dados

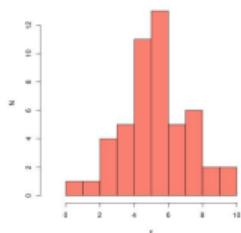
- dado um conjunto de dados, D , queremos *ajustar* um **modelo** M que depende de um certo número de parâmetros ajustáveis, w
- exemplo: ajuste de uma função $y = f(x; w)$ aos dados
 $M: y = a + bx \quad w = \{a, b\}$
- ajuste = *inferência* dos parâmetros do modelo e de seus erros



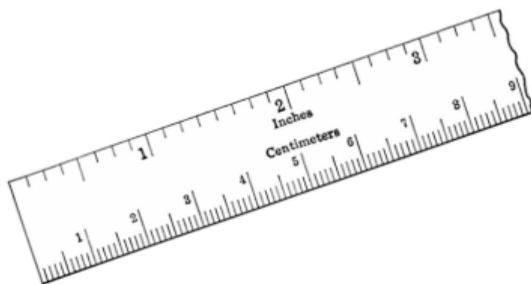
- modelos podem ser de natureza muito diversa:
 - funções simples para as quais o ajuste fornece w
ex.: ajuste de polinômios, gaussiana,...
 - funções teóricas onde conhecemos os parâmetros e queremos ver se eles ajustam os dados
ex.: a função de massa de aglomerados de galáxias com o modelo Λ CDM
 - funções muito complexas para as quais o ajuste fornece w , mas w não nos interessa: deep learning
 - ...

dados geralmente tem erros

- dois tipos de erros nos dados:
aleatórios e sistemáticos
 - suponha que você meça n vezes a distância entre dois pontos com uma régua
 - esperamos que a média e o desvio padrão dessas medidas sejam uma boa estimativa da distância entre os dois pontos e do erro da medida

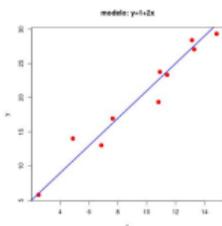


- mas e se você descobre que estava usando polegadas e pensava que estava usando centímetros?
(1 polegada = 2,54 cm)
Como isso afeta seus resultados?
- este é um exemplo de erros sistemáticos



dados geralmente tem erros

- erros nas medidas:
 - erros intrínsecos ao processo de medida decrescem com $1/\sqrt{n}$ e são em geral aleatórios
 - erros sistemáticos são introduzidos por processos que muitas vezes não se controla
 - ex.: ruído indesejado em um detector, problemas na sensibilidade dos filtros, estrelas de calibração erradas ...
 - o desafio é *identificá-los e corrigi-los* (modelá-los)
- erros homocedásticos: todas as medidas têm o mesmo erro
- erros heterocedásticos: as medidas têm erros diferentes
- **modelos também têm erros:**
erros epistêmicos
considere, por exemplo, o modelo que se obtém ajustando-se um polinômio de grau 10 aos dados da figura



a verossimilhança frequentista

- dado um conjunto de dados, D , queremos *ajustar* um **modelo** M que depende de um certo número de parâmetros ajustáveis, w
- $\mathcal{L}(w)$ - função de verossimilhança (ou verossimilhança):
probabilidade dos dados D , escrita como função dos parâmetros w
- exemplo- temos um conjunto de medidas de uma variável e o erro, $D = \{x_i, \sigma_i\}$, que queremos modelar como uma gaussiana de média w :

$$\mathcal{L}(w) \propto \exp\left(-\frac{\chi^2(w)}{2}\right), \quad \chi^2(w) = \sum_i \frac{(x_i - w)^2}{\sigma_i^2}$$

- R. A. Fisher (1912): *método da máxima verossimilhança*
- qual é a melhor estimativa para o parâmetro w de um modelo com base nos dados disponíveis?
a melhor estimativa é a que maximiza a verossimilhança

$$\left. \frac{d\mathcal{L}(w)}{dw} \right|_{MV} = 0, \quad \left. \frac{d^2\mathcal{L}(w)}{dw^2} \right|_{MV} < 0$$

(note que $\mathcal{L}(w)$ não é uma distribuição de probabilidades!)

exemplo: combinando medidas

- temos N medidas de uma quantidade x com erros heterocedásticos (sigmas diferentes): $D = \{x_i, \sigma_i\}$
- problema: como combinar as medidas para se obter uma melhor estimativa de x ?
- vamos supor que as duas medidas são independentes e extraídas de distribuições gaussianas onde o *valor verdadeiro* da quantidade medida é w :

$$P_i \propto \exp \left[-\frac{(x_i - w)^2}{2\sigma_i^2} \right]$$

- a verossimilhança será $\mathcal{L}(w) = \prod_i P_i \propto \exp \left(-\frac{\chi^2(w)}{2} \right)$, onde
$$\chi^2(w) = \sum_i \frac{(x_i - w)^2}{\sigma_i^2}$$
- é fácil ver que valor que maximiza a verossimilhança (ou que minimiza o χ^2) é

$$\hat{w} = \frac{\sum_i p_i x_i}{\sum_i p_i}, \quad \text{com } p_i = \frac{1}{\sigma_i^2}$$

- a melhor estimativa é a média ponderada das medidas, onde os pesos dependem do inverso das variâncias (= precisão)
- se os erros são iguais, \hat{w} é a média das medidas

exemplo: distribuição binomial

- joga-se uma moeda n vezes e obtém-se T caras e H coroas: ($n = T + H$) qual é a probabilidade de se obter uma coroa (w)?
- a verossimilhança é dada pela distribuição binomial com parâmetro w :

$$\mathcal{L}(w) = \binom{n}{H} w^H (1-w)^{n-H}$$

- estimativa de MV de w :

$$\frac{d \ln \mathcal{L}(w)}{dw} =$$

$$= \frac{d}{dw} \left[\ln \binom{n}{H} + H \ln w + (n-H) \ln(1-w) \right]$$

$$= \frac{H}{w} - \frac{n-H}{1-w} = 0$$

ou

$$\hat{w} = \frac{H}{n}$$

- a estimativa de MV de w é a fração observada de coroas, o que é intuitivo

exemplo: distribuição de Poisson

- dados D : observamos n eventos durante um intervalo de tempo t , $D = \{n, t\}$
- a probabilidade de se observar n eventos durante um intervalo de tempo t para uma taxa de eventos w é um processo poissoniano:

$$P(n, t|w) = \frac{(wt)^n}{n!} \exp(-wt)$$

- qual é a estimativa de MV para a taxa de eventos w ?

$$\mathcal{L}(w) \propto P(n, t|w)$$

$$\begin{aligned} \frac{d \ln \mathcal{L}(w)}{dw} &= \\ &= \frac{d}{dw} \left[n \ln(wt) - \ln n! - wt \right] = \end{aligned}$$

$$\frac{nt}{wt} - t = 0$$

ou

$$\hat{w} = \frac{n}{t}$$

- a estimativa de MV de w é a taxa média de eventos

exemplo: modelagem de uma função com erros gaussianos

- n dados $D = \{x_i, y_i, \sigma_i\}$, com erros σ_i , que queremos modelar com uma função com m parâmetros w : $y = f(x; w)$
- w : vetor de m parâmetros
- assumindo que as medidas são independentes e os erros gaussianos,

$$\begin{aligned}\mathcal{L}(w) &\propto \prod_{i=1}^n P(D_i|w) \\ &\propto \prod_{i=1}^n \exp \left[-\frac{(y_i - f(x_i; w))^2}{2\sigma_i^2} \right] \\ &\propto \exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(y_i - f(x_i; w))^2}{\sigma_i^2} \right]\end{aligned}$$

logo,

$$\mathcal{L}(w) \propto \exp \left[-\frac{1}{2} \chi^2(w) \right]$$

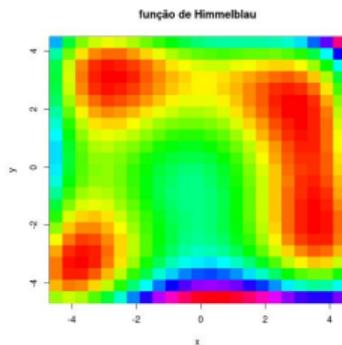
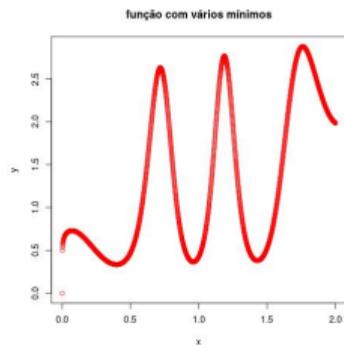
onde

$$\chi^2(w) = \sum_{i=1}^n \left[\frac{(y_i - f(x_i; w))^2}{\sigma_i^2} \right]$$

- solução de MV: minimização do $\chi^2(w)$
- \hat{w} pode, em alguns casos, ser obtido analiticamente ou, na maioria dos casos, numericamente
- boas soluções têm $\chi_{red}^2 = \chi^2 / (m - 1) \simeq 1$
(χ_{red}^2 : " χ^2 reduzido")

máxima verossimilhança como otimização de funções

- minimização é um problema de *otimização de funções*
- mínimos: podem ser *locais* ou o mínimo *global*
- o mínimo global, no caso geral, pode ser difícil de se achar

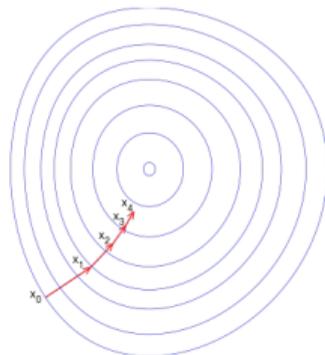


- algoritmo do “*gradiente descendente*”
procedimento iterativo:

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \gamma \nabla \chi^2(\mathbf{w}) \quad (0 < \gamma < 1)$$

sempre acha um mínimo (local)

- γ é chamado de *taxa de aprendizado*



erro nos parâmetros de um modelo

- vamos considerar um modelo com um único parâmetro, w
- expandindo $\mathcal{L}(w)$ em série de Taylor até segunda ordem em torno do máximo \hat{w} :

$$\ln \mathcal{L}(w) = \ln \mathcal{L}(\hat{w}) + \left. \frac{d \ln \mathcal{L}(w)}{dw} \right|_{MV} (w - \hat{w}) + \frac{1}{2} \left. \frac{d^2 \ln \mathcal{L}(w)}{dw^2} \right|_{MV} (w - \hat{w})^2 + \dots$$

- como o segundo termo se anula na vizinhança de \hat{w} :

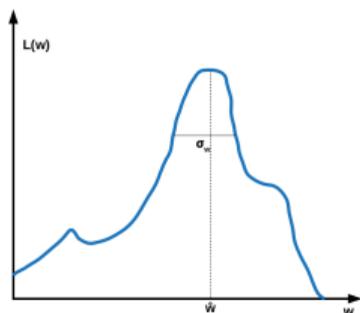
$$\mathcal{L}(w) \approx \mathcal{L}(\hat{w}) \exp \left[- \frac{(w - \hat{w})^2}{2\sigma_w^2} \right] \quad \text{onde} \quad \frac{1}{\sigma_w^2} = - \left. \frac{d^2 \ln \mathcal{L}(w)}{dw^2} \right|_{MV}$$

- a verossimilhança em torno do máximo pode ser aproximada por uma gaussiana de largura σ_w

- σ_w é o erro atribuído a \hat{w}
- exemplo: combinação de duas medidas

$$\sigma_w^2 = \frac{1}{p_A + p_B}$$

onde $p_A = 1/\sigma_A^2$ e $p_B = 1/\sigma_B^2$



erro nos parâmetros de um modelo

- no caso de vários parâmetros $\mathbf{w} = \{w_1, w_2, \dots, w_m\}$, \mathcal{L} em torno do máximo pode ser aproximada por uma gaussiana multivariada:

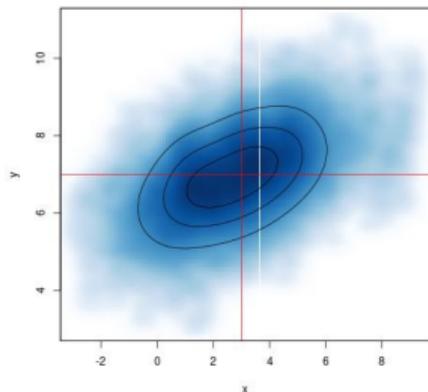
$$\mathcal{L}(\mathbf{w}) \approx \mathcal{L}(\mathbf{w}_{MV}) \exp \left[-\frac{1}{2}(\mathbf{w} - \mathbf{w}_{MV})\mathbf{C}^{-1}(\mathbf{w} - \mathbf{w}_{MV})^T \right] + \dots$$

\mathbf{C} : matriz de covariância ($m \times m$)

- hessiano* \mathbf{H} : matriz com as derivadas segundas de \mathcal{L}
- matriz de informação de Fisher*: negativo do valor esperado do hessiano no máximo:

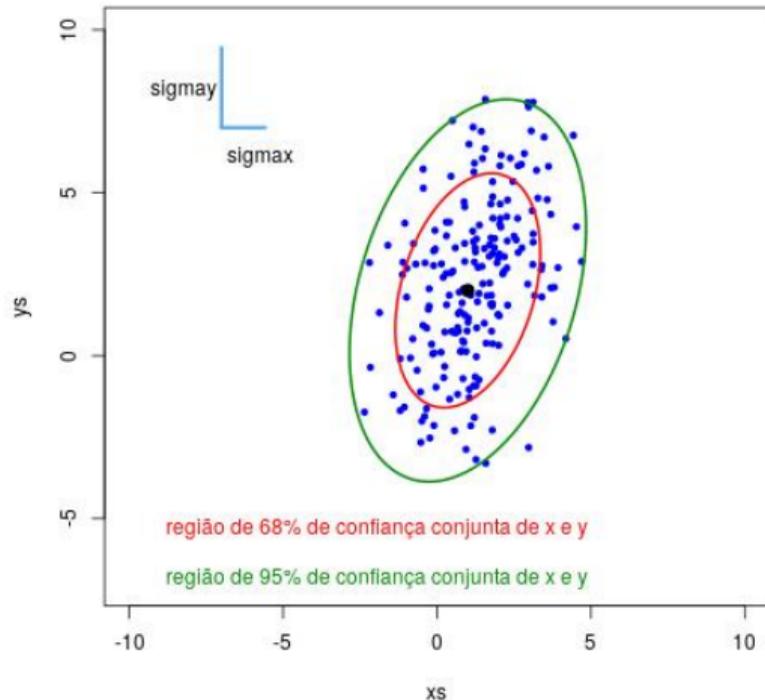
$$I_{jk} = -E(H_{jk})_{MV} = -\left. \frac{\partial^2 \ln \mathcal{L}(\mathbf{w})}{\partial w_j \partial w_k} \right|_{MV}$$

- matriz de covariância*: $\mathbf{C} = \mathbf{I}^{-1}$ (inversa da matriz de informação)
- erros nos parâmetros: $\sigma_j^2 = C_{jj}$



erro nos parâmetros de um modelo

- note que nem sempre σ é uma boa forma de representar as incertezas em um ajuste
- quando se tem vários parâmetros, elipses de erro (considerando os parâmetros dois a dois) podem ser mais informativas
- *bootstrap* é um jeito fácil de estimar esse tipo de incerteza



propagação de erros:

- vamos supor que temos uma quantidade f que é função de ao menos duas variáveis, x_1 e x_2 : $f = f(x_1, x_2, \dots)$
- se medimos x_1 e x_2 , como o erro nessas variáveis, σ_{x_1} e σ_{x_2} , afeta f , supondo que as medidas e seus erros são independentes?
- considerando pequenos desvios δ em torno do ponto (x_1, x_2, \dots) , expandindo f em série de Taylor e retendo apenas os termos de primeira ordem da expansão:

$$\delta f = \left(\frac{\partial f}{\partial x_1} \right)_{x_1} \delta x_1 + \left(\frac{\partial f}{\partial x_2} \right)_{x_2} \delta x_2 + \dots$$

- como a variância de uma soma é a soma das variâncias, a variância esperada de f será:

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x_1} \right)_{x_1}^2 \sigma_{x_1}^2 + \left(\frac{\partial f}{\partial x_2} \right)_{x_2}^2 \sigma_{x_2}^2 + 2 \left(\frac{\partial f}{\partial x_1} \right)_{x_1} \left(\frac{\partial f}{\partial x_2} \right)_{x_2} \rho \sigma_{x_1} \sigma_{x_2} \dots$$

ajuste de uma reta aos dados

- suponha que temos n dados $\{x_i, y_i\}$, com erros σ em y_i
- queremos modelar os dados com uma reta:
 $y = f(x; a, b) = a + bx$
- este problema é denominado *regressão linear ordinária (OLS)*
 - regressão: procedimentos para estimar relações entre variáveis
 y é uma variável contínua
(se y é discreta: *classificação*)
 - linear: *em relação aos parâmetros a, b*
- verossimilhança de uma medida:
$$\mathcal{L}_i(a, b) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(y_i - a - bx_i)^2}{2\sigma^2} \right]$$

- verossimilhança da amostra:

$$\mathcal{L}(a, b) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(y_i - a - bx_i)^2}{2\sigma^2} \right] = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left[-\frac{1}{2}\chi^2 \right]$$

- onde

$$\chi^2(a, b) = \sum_{i=1}^n \frac{(y_i - a - bx_i)^2}{\sigma^2}$$

- $\{a, b\}$ - obtidos pela maximização de \mathcal{L} ou pela minimização do χ^2

ajuste de uma reta aos dados

- minimização do χ^2 :

$$\frac{d\chi^2}{da} = -\frac{2}{\sigma} \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\frac{d\chi^2}{db} = -\frac{2}{\sigma} \sum_{i=1}^n x_i (y_i - a - bx_i) = 0$$

- que podemos reescrever como um sistema de equações simultâneas para a e b :

$$\begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

- cuja solução é

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \frac{\sum x_i^2}{n \sum x_i^2 - (\sum x_i)^2} & \frac{-\sum x_i}{n \sum x_i^2 - (\sum x_i)^2} \\ \frac{-\sum x_i}{n \sum x_i^2 - (\sum x_i)^2} & \frac{n}{n \sum x_i^2 - (\sum x_i)^2} \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

- ou

$$a = \frac{\sum x_i^2}{n \sum x_i^2 - (\sum x_i)^2} \sum y_i + \frac{-\sum x_i}{n \sum x_i^2 - (\sum x_i)^2} \sum x_i y_i$$

$$b = \frac{-\sum x_i}{n \sum x_i^2 - (\sum x_i)^2} \sum y_i + \frac{n}{n \sum x_i^2 - (\sum x_i)^2} \sum x_i y_i$$

- matriz de covariância:

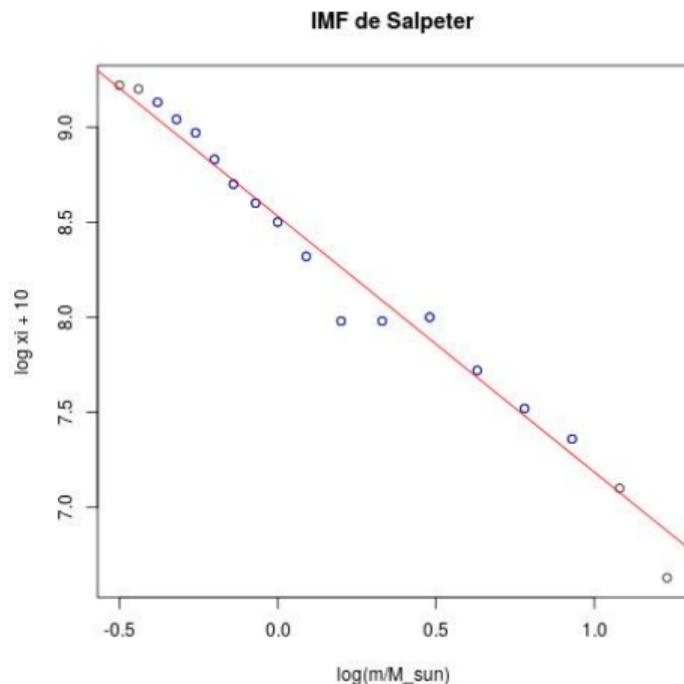
$$C = \sigma^2 \begin{bmatrix} \frac{\sum x_i^2}{n \sum x_i^2 - (\sum x_i)^2} & \frac{-\sum x_i}{n \sum x_i^2 - (\sum x_i)^2} \\ \frac{-\sum x_i}{n \sum x_i^2 - (\sum x_i)^2} & \frac{n}{n \sum x_i^2 - (\sum x_i)^2} \end{bmatrix} = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}$$

extensões simples da regressão linear

note que muitas vezes se pode fazer mudanças de variáveis e aplicar a regressão linear:

- $y = A10^{bx} \rightarrow \log y = \log A + bx$
- $y = Ax^b \rightarrow \log y = \log A + b \log x$

- $f(y) = a + bg(x)$
 - f e g podem ser funções arbitrárias!
 - o importante é que o **modelo seja linear nos parâmetros** a e b (i.e., não dependa de a^2, ab, \dots)



modelos lineares

- exemplos de modelos lineares nos parâmetros w

- $y = w_0 + w_1x$
- $y = w_0 + w_1x + w_2x^2 + w_3x^3$
- $y = w_0e^x + w_1\sin(x)$

- exemplos de modelos não-lineares nos parâmetros

- $y = w_0/(1 + w_1x)^2$
- $y = w_0\sin(x + w_1) + w_2\cos(x + w_3)$
- $y = w_0e^{-w_1x}$
(mas note que $\ln y$ é linear)

mínimos quadrados linear geral

- modelos lineares são muito úteis, pois podem envolver funções arbitrárias de x
- modelo linear com m parâmetros a_k :

$$y = \sum_k a_k X_k(x),$$

onde $X_k(x)$ são m funções arbitrárias de x

- exemplo: polinômios de grau n :
 $y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$

- máxima verossimilhança = minimização do χ^2 :

$$\chi^2(\mathbf{a}) = \sum_{i=1}^n \left[\frac{y_i - \sum_k a_k X_k(x_i)}{\sigma_i} \right]^2$$

- solução:

$$a_j = \sum_{k=1}^m [\alpha]_{jk}^{-1} \beta_k = \sum_{k=1}^m C_{jk} \beta_k,$$

$$\alpha_{kj} = \sum_{i=1}^n \frac{X_j(x_i) X_k(x_i)}{\sigma_i^2} \quad \beta_k = \sum_{i=1}^n \frac{y_i X_k(x_i)}{\sigma_i^2}$$

- $\mathbf{C} = \alpha^{-1}$: matriz de covariância
- erros dos parâmetros: $\sigma(a_j)^2 = C_{jj}$

modelos não-lineares

- em geral, em problemas não-lineares, os parâmetros que minimizam o $\chi^2(\mathbf{w})$

$$\chi^2(\mathbf{w}) = \sum_{i=1}^N \left[\frac{y_i - f(x_i; \mathbf{w})}{\sigma_i} \right]^2$$

devem ser obtidos numericamente

- exemplo: ajuste do perfil de brilho da galáxia NGC 4472 (M49) com a lei de Sérsic:

$$\log I(r) = \log I_e - b_n [(r/r_e)^{1/n} - 1]$$

$$b_n \simeq 0.868n - 0.142, \text{ para } 0.5 < n < 16.5$$

- 3 parâmetros: $\mathbf{w} = \{I_e, r_e, n\}$
 - r_e : raio efetivo - raio que contém metade da luminosidade da galáxia
 - I_e : brilho superficial em r_e
 - n : expoente de Sérsic
- o perfil de brilho superficial é medido em unidades de mag/arcsec^2 :

$$\mu(r) = \mu_0 - 2.5 \log I(r) \text{ mag/arcsec}^2$$

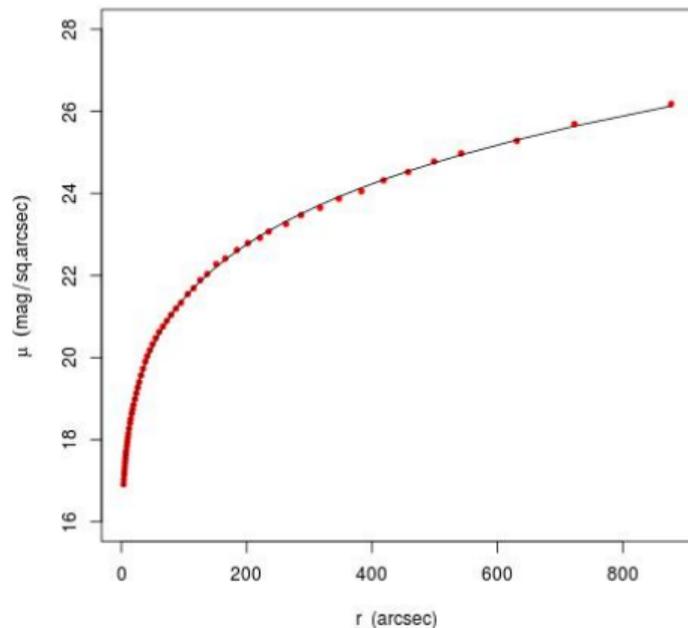
- logo, se $\mu'_0 = \mu_0 - 2.5 \log I_e$,

$$\mu(r) = \mu'_0 + 2.5 b_n [(r/r_e)^{1/n} - 1]$$

modelos não-lineares



M49: galáxia mais luminosa do aglomerado de Virgo



exemplo: perfil de brilho de NGC 4472 (M49)

- 1000 simulações de bootstrap do ajuste
- note que os parâmetros são correlacionados

