AGA 0505 - Análise de Dados em Astronomia

6. O Método da Máxima Verossimilhança

Laerte Sodré Jr.

1o. semestre, 2023

aula de hoje:

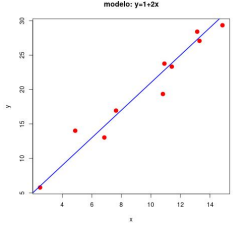
1. modelagem dos dados
2. a verossimilhança
3. inferência de parâmetros pelo método da máxima verossimilhança
4. exemplo: combinação de medidas
5. exemplo: distribuição binomial
6. exemplo: distribuição poissoniana
7. exemplo: ajuste de funções com erros gaussianos
8. otimização de funções
9. incertezas 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: $y = a + bx \qquad w = \{a, b\}$
- ajuste $=$ *inferência* dos parâmetros do modelo



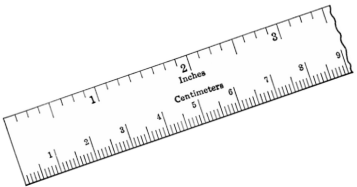
- modelos podem ser de natureza muito diversa:
  - funções simples para as quais o ajuste fornece w
    ex.: ajuste de polinômios, gaussiana,...
  - funções teóricas onde conhecemos os parâmetros e queremos ver se eles ajustam os dados
    ex.: a função de massa de aglomerados de galáxias com o modelo ΛCDM
  - 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
  - erros intrínsecos ao processo de medida decrescem como $1/\sqrt{n}$ e são em geral aleatórios

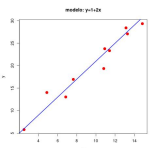
- mas e se você descobre que estava usando polegadas e pensava que estava usando centímetros?
  (1 polegada $=$ 2,54 cm)
  Como isso afeta seus resultados?

- 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



modelagem probabilística dos dados: a verossimilhança

- dado um conjunto de dados, D, queremos *ajustar* um **modelo** M que depende de um certo número de parâmetros ajustáveis, w
- inferência bayesiana de w:
  estimativa do posterior $P(w|D,M)$

$$P(w|D,M) \propto P(D|w,M)P(w|M)$$

  $P(D|w,M)$: verossimilhança
- a verossimilhança bayesiana é uma função dos dados, D
- a mesma função frequentista é vista como função dos parâmetros: $\mathcal{L}(w)$
- $\mathcal{L}(w)$ não é uma PDF!

- exemplo: temos um conjunto de medidas de uma variável, $D = \{x_1, ...x_n\}$ que queremos modelar como uma gaussiana de parâmetros $w = (\mu, \sigma)$
  - verossimilhança bayesiana:
    $P(D|w,M) \propto \exp\left(-\frac{\chi^2(D)}{2}\right)$
  - verossimilhança frequentista:
    $\mathcal{L}(w) \propto \frac{1}{\sigma}\exp\left(-\frac{\chi^2(w)}{2}\right)$
  - nos dois casos,

$$\chi^2 = \sum_i \frac{\left(x_i - \mu\right)^2}{\sigma^2}$$

modelagem dos dados e verossimilhança

- modelagem frequentista:
  - 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
  - *intervalo de confiança*: quão confiante podemos estar no valor estimado do parâmetro?
    o intervalo de confiança dá o intervalo em que se espera encontrar o parâmetro com um certo nível de confiança (p.ex., 95%)

- verossimilhança frequentista:
  $\mathcal{L}(w) \propto P(D|w, M)$
  - é função dos parâmetros: $\mathcal{L}(w)$
  - NÃO É uma PDF: não é ou precisa ser normalizada
  - os parâmetros são estimados maximizando-se $\mathcal{L}(w)$ (ou $\log \mathcal{L}$)

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

exemplo: combinando medidas

- temos duas medidas independentes de uma quantidade x com erros heterocedásticos (sigmas diferentes):

$$x = x_A \pm \sigma_A \qquad x = x_B \pm \sigma_B$$

- problema: como combinar as medidas A e B para se obter uma melhor estimativa de x?
- alerta: note que se $|x_A - x_B|$ for muito maior que os desvios padrão, os dados podem estar sendo afetados por erros de outra natureza (dados *inconsistentes: outliers*)
- vamos resolver este problema pelo método da Máxima Verossimilhança

- vamos supor que as duas medidas são independentes e extraídas de distribuições gaussianas onde o valor verdadeiro da quantidade medida é μ:

$$P(x_A, \sigma_A|\mu) \propto \frac{1}{\sigma_A} \exp\left[-\frac{(x_A-\mu)^2}{2\sigma_A^2}\right]$$

$$P(x_B, \sigma_B|\mu) \propto \frac{1}{\sigma_B} \exp\left[-\frac{(x_B-\mu)^2}{2\sigma_B^2}\right]$$

e nosso objetivo é estimar μ

exemplo: combinando medidas

- a verossimilhança $\mathcal{L}(\mu)$ é proporcional à probabilidade conjunta de x_A e x_B
- supondo que as medidas são independentes:

$$\mathcal{L}(\mu) \propto P(x_A, x_B, \sigma_A, \sigma_B|\mu) =$$

$$P(x_A, \sigma_A|\mu)P(x_B, \sigma_B|\mu) \propto$$

$$\exp\left(-\frac{\chi^2}{2}\right)$$

onde

$$\chi^2(\mu) = \left(\frac{x_A-\mu}{\sigma_A}\right)^2 + \left(\frac{x_B-\mu}{\sigma_B}\right)^2$$

- $\hat{\mu}$: valor que maximiza a verossimilhança (ou que minimiza o χ^2)
- é fácil ver que

$$\hat{\mu} = \frac{w_A x_A + w_B x_B}{w_A + w_B}$$

onde $w_A = 1/\sigma_A^2$ e $w_B = 1/\sigma_B^2$

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

exemplo: distribuição binomial

- joga-se uma moeda n vezes e obtém-se T caras e H coroas:
  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:

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

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

$$\mathcal{L}(w) \propto P(D|w)$$

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

$$= \frac{d}{dw}\left[n \ln(wt) - \ln n! - wt\right] =$$

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

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,

$$\mathcal{L}(w) \propto P(D|w) = \prod_{i=1}^{n} P(D_i|w)$$

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

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

logo,

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

onde

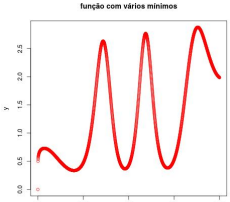
$$\chi^2(w) = \sum_{i=1}^{n} \left[\frac{(y_i - f(x_i; w))^2}{\sigma_i^2} \right]$$

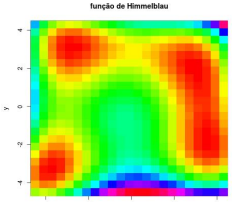
- solução de MV: minimização do $\chi^2(w)$
- $\hat{w}$ pode, em alguns casos, ser obtido analiticamente ou, na maioria dos casos, numericamente
- boas soluções têm $\chi^2_{red} = \chi^2/(m-1) \simeq 1$

(χ^2_{red}: "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



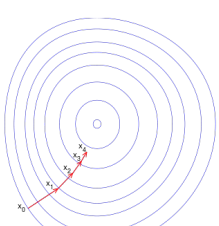


- algoritmo do *"gradiente descendente"* procedimento iterativo:

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

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 em torno do máximo $\hat{w}$:

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

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

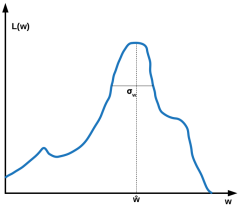
$$\mathcal{L}(w) \approx \mathcal{L}(\hat{w}) \exp\left[-\frac{(w-\hat{w})^2}{2\sigma_w^2}\right] + \ldots \quad \text{onde} \quad \frac{1}{\sigma_w^2} = -\left.\frac{d^2 \ln \mathcal{L}(w)}{dw^2}\right|_{MV}$$

- 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

$$\sigma_w^2 = \frac{1}{w_A + w_B}$$

onde $w_A = 1/\sigma_A^2$ e $w_B = 1/\sigma_B^2$



erro nos parâmetros de um modelo

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

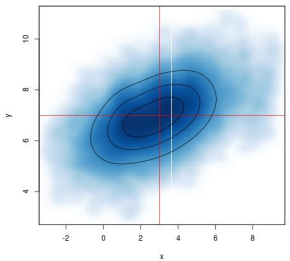
$$\mathcal{L}(\mathbf{w}) \approx \mathcal{L}(\mathbf{w}_{MV}) \exp\left[-\frac{1}{2}(\mathbf{w} - \mathbf{w}_{MV})\mathbf{C}^{-1}(\mathbf{w} - \mathbf{w}_{MV})T \right] + ...$$

$\mathbf{C}$: 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:

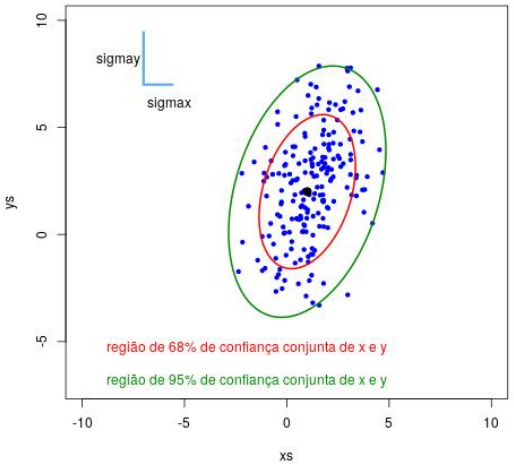
$$I_{jk} = -E(H_{jk})_{MV} = -\left.\frac{\partial^2 \ln \mathcal{L}(\mathbf{w})}{\partial w_j \partial w_k}\right|_{MV}$$

- *matriz de covariância*: $C = I^{-1}$ (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
- receita para determinar elipses de erro em pares de parâmetros:
  Coe, arXiv:0906.4123



propagação de erros:

- vamos supor que temos uma quantidade f que é função de ao menos duas variáveis, x_1 e x_2: $f = f(x_1, x_2, ...)$
- 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 $(x_1, x_2, ...)$, expandindo f em série de Taylor e retendo apenas os termos de primeira ordem da expansão:

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

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

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

ajuste de uma reta aos dados

- suponha que temos n dados $\{x_i, y_i\}$ que queremos modelar com uma reta: $y = f(x; a, b) = a + bx$

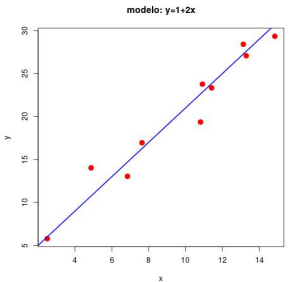
- este problema é denominado *regressão linear*
  - regressão: procedimentos para estimar relações entre variáveis
    y é uma variável contínua
    (se y é discreta: *classificação*)
  - linear: *em relação aos parâmetros* a,b

- supomos que conhecemos os erros nas medidas σ_i e que os x_i são conhecidos exatamente

- nesse caso,

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

- a, b - obtidos pela minimização do χ^2



ajuste de uma reta aos dados

- equação da reta: $y = a + bx$
- χ^2:

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

- a, b - minimização do χ^2:

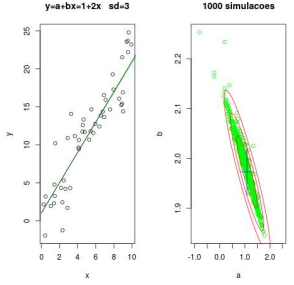
$$a = \frac{S_{xx}S_y - S_x S_{xy}}{\Delta} \qquad b = \frac{SS_{xy} - S_x S_y}{\Delta}$$

$$S = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \qquad S_x = \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \qquad S_y = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}$$

$$S_{xx} = \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} \qquad S_{xy} = \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2} \qquad \Delta = SS_{xx} - S_x^2$$

- a variância dos parâmetros, a e b, pode ser obtida via $\mathbf{C}^{-1}$ ou bootstrap:

$$\sigma_a^2 = \frac{S_{xx}}{\Delta} \qquad \sigma_b^2 = \frac{S}{\Delta}$$

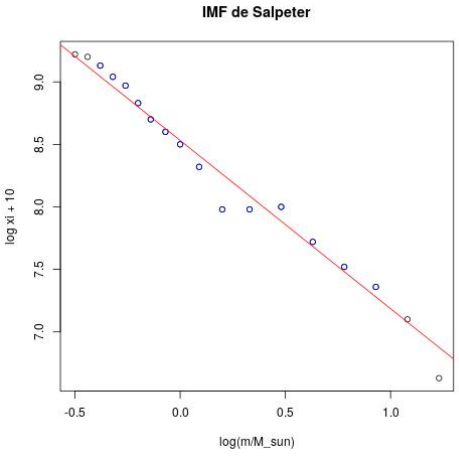


extensões simples da regressão linear

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

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

- $f(y) = a + bg(x)$
  - 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 $a^2, ab, ...$)



modelos lineares

- exemplos de modelos lineares nos parâmetros w
  - $y = w_0 + w_1 x$
  - $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$
  - $y = w_0 e^x + w_1 \sin(x)$

- exemplos de modelos não-lineares nos parâmetros
  - $y = w_0/(1 + w_1 x)^2$
  - $y = w_0 \sin(x + w_1) + w_2 \cos(x + w_3)$
  - $y = w_0 e^{-w_1 x}$
    (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: $y = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$

- 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:

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

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

- $\mathbf{C} = \alpha^{-1}$: 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 de NGC 4472 (M49) com a lei de Sérsic:

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

3 parâmetros: $\mathbf{w} = \{I_e, r_e, n\}$ $\quad (b_n = f(n))$

M49: galáxia mais luminosa do aglomerado de Virgo

exemplo: perfil de brilho de NGC 4472 (M49)

- o perfil de brilho das galáxias é frequentemente modelado por um *perfil de Sérsic:*

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

$(b_n \simeq 0.868n - 0.142, \textit{para}\, 0.5 < n < 16.5)$

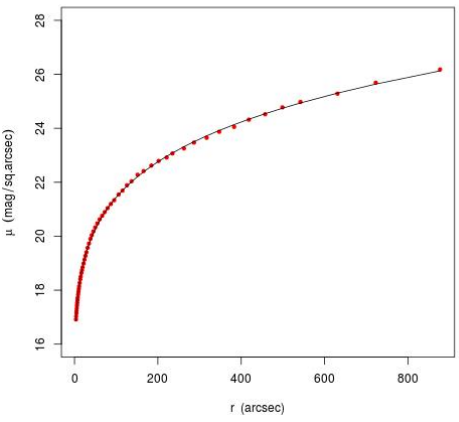
  - *o perfil de brilho superficial é medido em unidades de mag/arcsec^2:*

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

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

$$\mu(r) = \mu_0' + 2.5 b_n[(r/r_e)^{1/n} - 1]$$

  - *parâmetros:* $\mathbf{w} = \{\mu_0', n, r_e\}$



exemplo: perfil de brilho de NGC 4472 (M49)

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