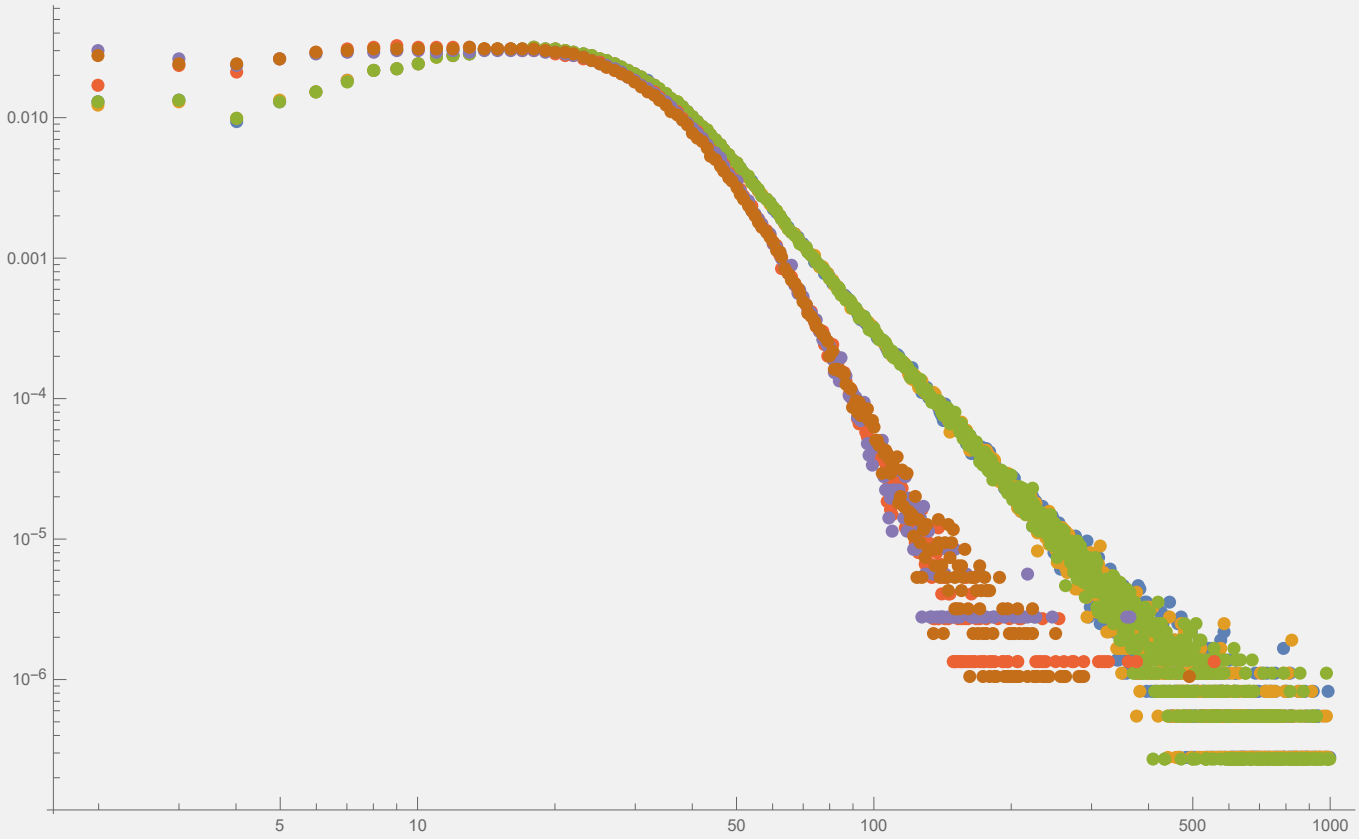
Bayesi döntéshozatal nemegyenlő tartójú eloszlásokra

HLT szeminárium

Borbély Gábor

2018.02.09.

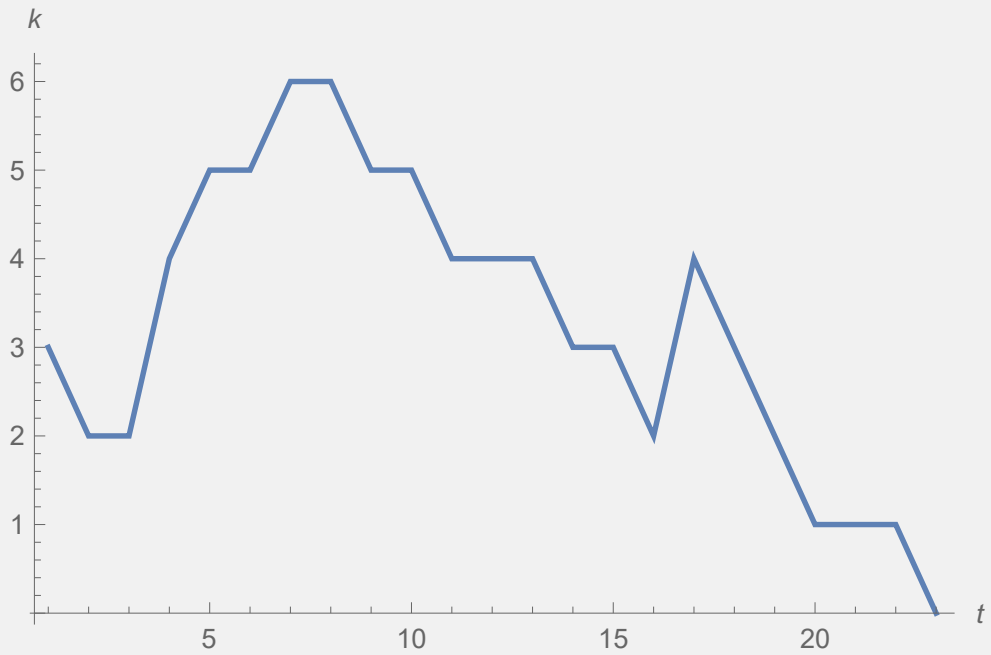
Adat



ábra : UKWAC és BNC (kisebb) korpuszok mondathossz eloszlása

Modell – I.

- Valency (“vegyérték”)



ábra : Random walk visszatérési idő, $k = 3$

Modell – II.

- $X_{t+1} = X_t + \eta$

Modell – II.

- $X_{t+1} = X_t + \eta$
- $\eta = -1, 0, 1, 2$ valamilyen p_{-1}, p_0, p_1 és p_2 valószínűségekkel

Modell – II.

- $X_{t+1} = X_t + \eta$
- $\eta = -1, 0, 1, 2$ valamilyen p_{-1}, p_0, p_1 és p_2 valószínűségekkel
  - $1, 2$: ige egy vagy két vonzattal

Modell – II.

- $X_{t+1} = X_t + \eta$
- $\eta = -1, 0, 1, 2$ valamilyen p_{-1}, p_0, p_1 és p_2 valószínűségekkel
  - $1, 2$: ige egy vagy két vonzattal
  - 0: például jelző

Modell – II.

- $X_{t+1} = X_t + \eta$
- $\eta = -1, 0, 1, 2$ valamilyen p_{-1}, p_0, p_1 és p_2 valószínűségekkel
  - $1, 2$: ige egy vagy két vonzattal
  - 0: például jelző
  - -1: főnév

Modell – II.

- $X_{t+1} = X_t + \eta$
- $\eta = -1, 0, 1, 2$ valamilyen p_{-1}, p_0, p_1 és p_2 valószínűségekkel
  - $1, 2$: ige egy vagy két vonzattal
  - 0: például jelző
  - -1: főnév
- k a kezdeti vegyérték

Modell – II.

- $X_{t+1} = X_t + \eta$
- $\eta = -1, 0, 1, 2$ valamilyen p_{-1}, p_0, p_1 és p_2 valószínűségekkel
  - $1, 2$: ige egy vagy két vonzattal
  - 0: például jelző
  - -1: főnév
- k a kezdeti vegyérték
- τ_k legyen az az idő amikor a k értékről indított folyamat először eléri a 0-t.

Modell – II.

- $X_{t+1} = X_t + \eta$
- $\eta = -1, 0, 1, 2$ valamilyen p_{-1}, p_0, p_1 és p_2 valószínűségekkel
  - $1, 2$: ige egy vagy két vonzattal
  - 0: például jelző
  - -1: főnév
- k a kezdeti vegyérték
- τ_k legyen az az idő amikor a k értékről indított folyamat először eléri a 0-t.
- τ_k nem más, mint k darab független τ_1 összege

Modell – II.

- $X_{t+1} = X_t + \eta$
- $\eta = -1, 0, 1, 2$ valamilyen p_{-1}, p_0, p_1 és p_2 valószínűségekkel
  - $1, 2$: ige egy vagy két vonzattal
  - 0: például jelző
  - -1: főnév
- k a kezdeti vegyérték
- τ_k legyen az az idő amikor a k értékről indított folyamat először eléri a 0-t.
- τ_k nem más, mint k darab független τ_1 összege
- τ_k eloszlás kell a $k, p_{-1}, \ldots p_2$ paraméterek függvényében.

Modell – III.

- τ_k eloszlásának meghatározása generátorfüggvény módszerrel

Modell – III.

- τ_k eloszlásának meghatározása generátorfüggvény módszerrel
- Lagrange inverziós tétele

Modell – III.

- τ_k eloszlásának meghatározása generátorfüggvény módszerrel
- Lagrange inverziós tétele
-

$$\mathbb{P}(\tau_k = n) = \frac{k}{n}[u^{n-k}]F^n(u)$$

Modell – III.

- τ_k eloszlásának meghatározása generátorfüggvény módszerrel
- Lagrange inverziós tétele
- $$\mathbb{P}(\tau_k = n) = \frac{k}{n}[u^{n-k}]F^n(u)$$
- ez utóbbi már $\mathcal{O}(n \cdot P(deg(F)))$ lépésben számolható

Modell – III.

- τ_k eloszlásának meghatározása generátorfüggvény módszerrel
- Lagrange inverziós tétele
- $$\mathbb{P}(\tau_k = n) = \frac{k}{n}[u^{n-k}]F^n(u)$$
- ez utóbbi már $\mathcal{O}(n \cdot P(deg(F)))$ lépésben számolható
- akárhányat léphet “felfelé” a folyamat ($deg(F)$)

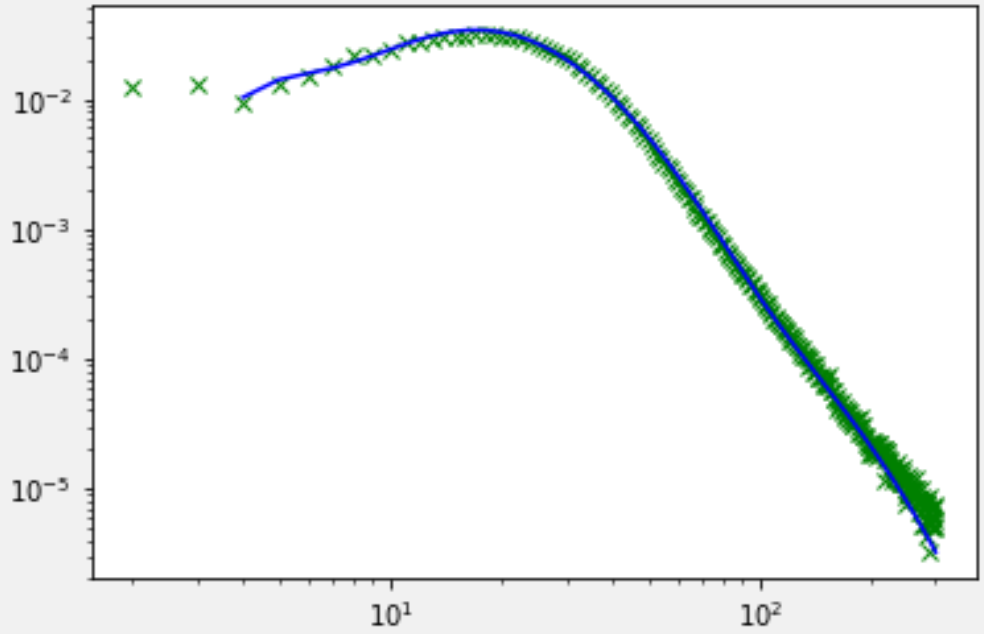
Modell – III.

- τ_k eloszlásának meghatározása generátorfüggvény módszerrel
- Lagrange inverziós tétele
- $$\mathbb{P}(\tau_k = n) = \frac{k}{n}[u^{n-k}]F^n(u)$$
- ez utóbbi már $\mathcal{O}(n \cdot P(deg(F)))$ lépésben számolható
- akárhányat léphet “felfelé” a folyamat ($deg(F)$)
- adott n-re és k-ra differenciálható függvénye a paramétereknek

Modell – III.

- τ_k eloszlásának meghatározása generátorfüggvény módszerrel
- Lagrange inverziós tétele
- $$\mathbb{P}(\tau_k = n) = \frac{k}{n}[u^{n-k}]F^n(u)$$
- ez utóbbi már $\mathcal{O}(n \cdot P(deg(F)))$ lépésben számolható
- akárhányat léphet “felfelé” a folyamat ($deg(F)$)
- adott n-re és k-ra differenciálható függvénye a paramétereknek
- Mixture modell, több k konvex lineáris kombinációja

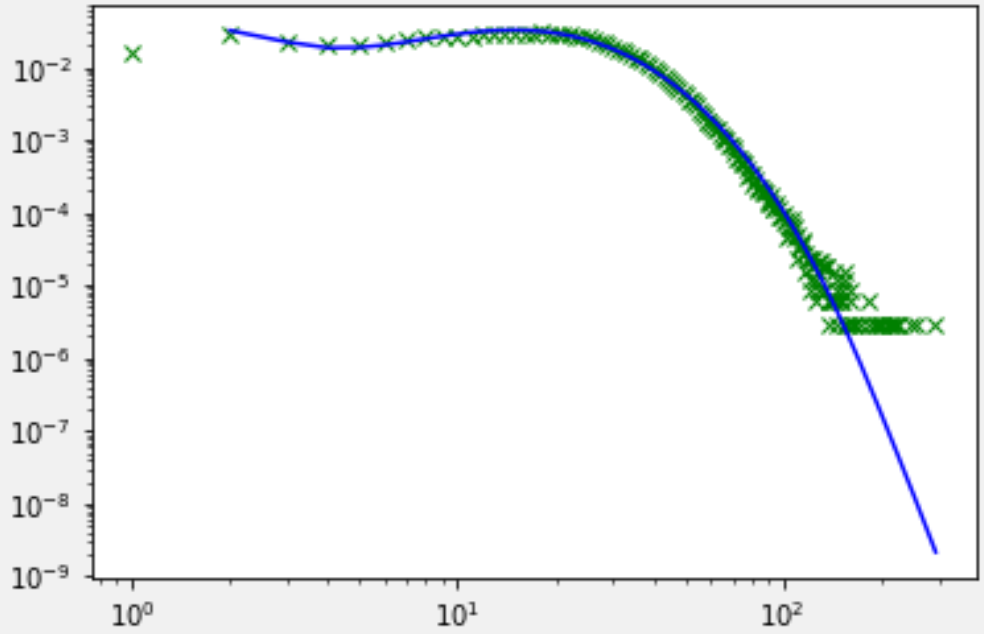
Illesztett paraméterek



ábra : UKWAC-20

k	α	m	p_{-1}	p_0	p_1	p_2
4	0.255934	-0.15220913291	0.447826	0.330732	0.147268	0.074174
5	0.744066	-0.184783674777	0.230145	0.733849	0.0266519	0.00935459

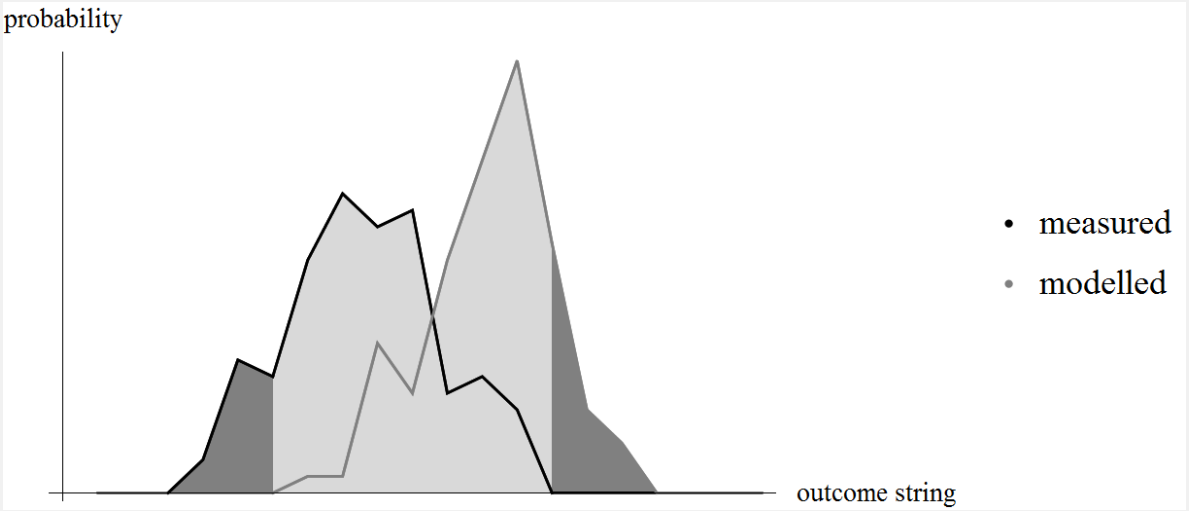
Illesztett paraméterek



ábra : BNC-E

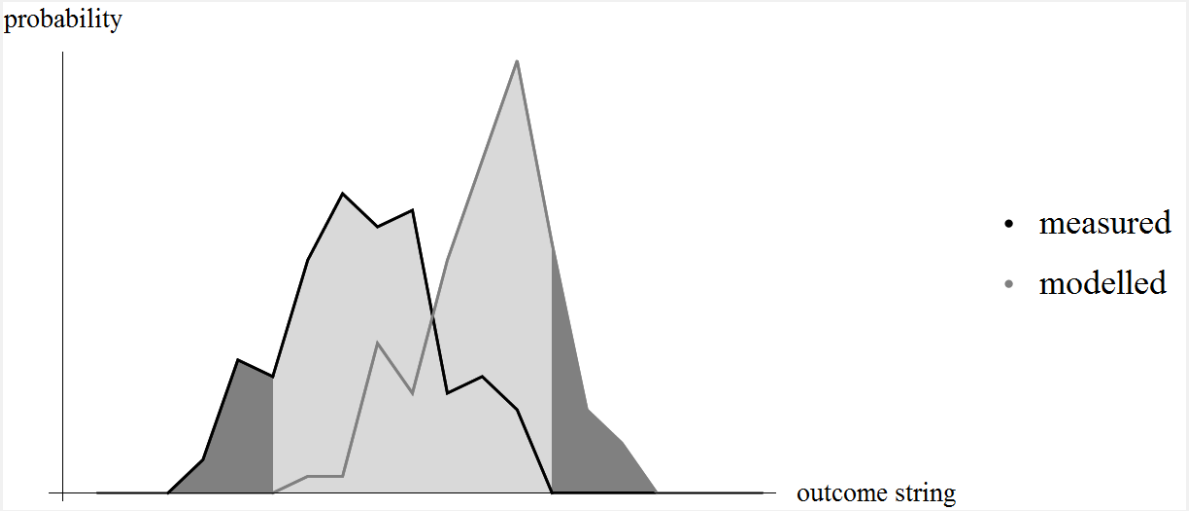
k	α	m	p_{-1}	p_0	p_1	p_2
2	0.12882	-0.243815563619	0.497859	0.315759	0.118719	0.0676624
3	0.87118	-0.12110476708	0.131467	0.860311	0.00608112	0.0021406

Általános probléma
nemegyenlő tartójú eloszlások összehasonlítása



ábra : nemegyenlő tartójú eloszlások

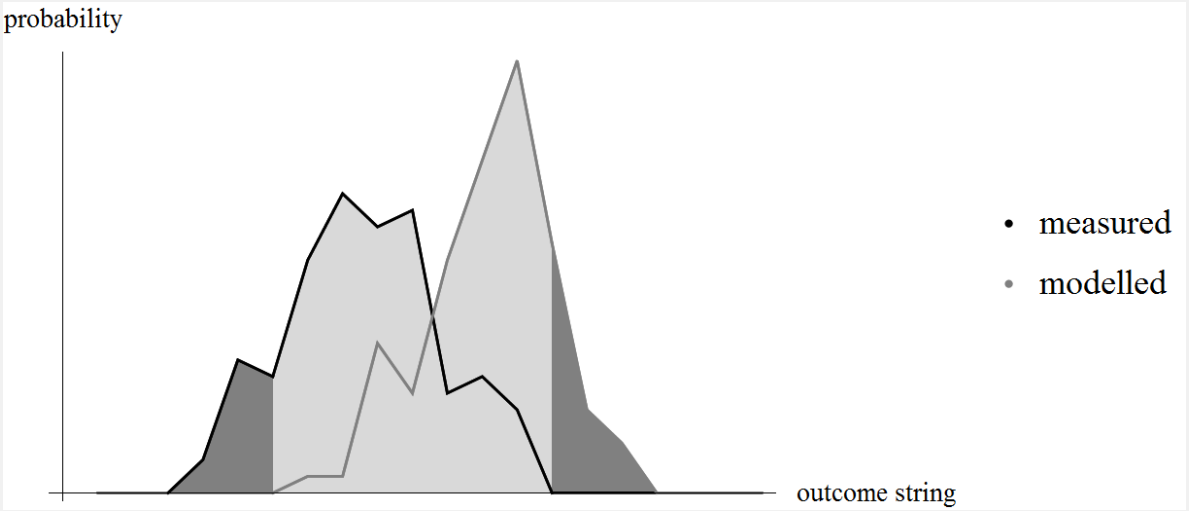
Általános probléma
nemegyenlő tartójú eloszlások összehasonlítása



ábra : nemegyenlő tartójú eloszlások

1. “Measured but not modeled”

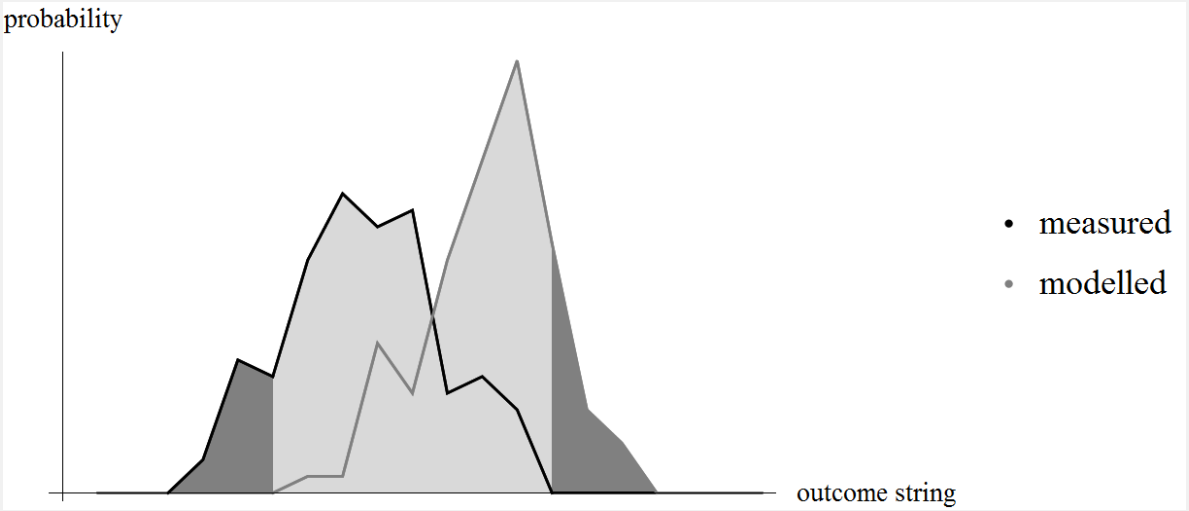
Általános probléma
nemegyenlő tartójú eloszlások összehasonlítása



ábra : nemegyenlő tartójú eloszlások

1. “Measured but not modeled”
   - dummy modell a lefedetlen részre

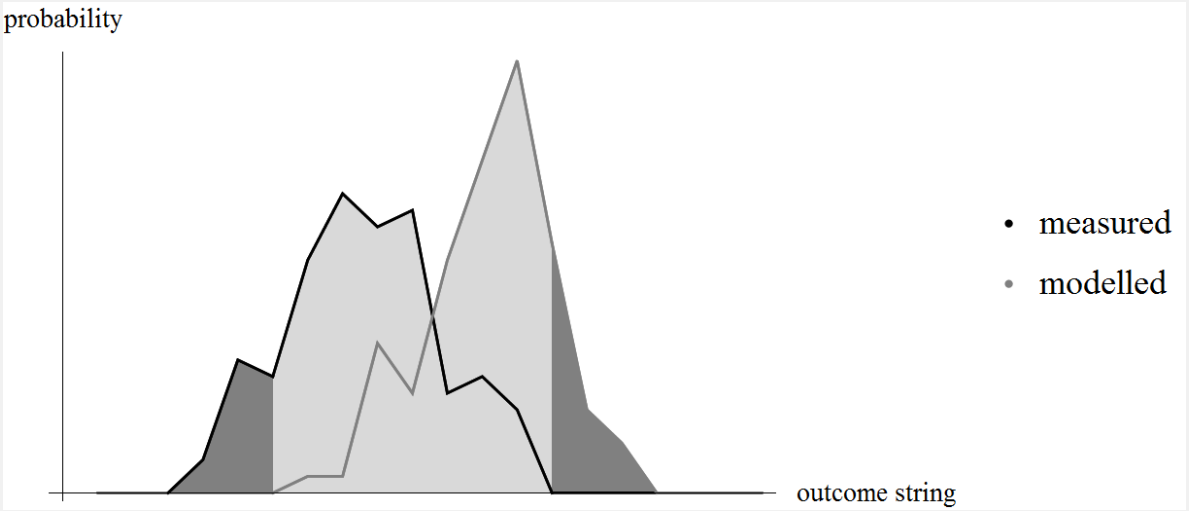
Általános probléma
nemegyenlő tartójú eloszlások összehasonlítása



ábra : nemegyenlő tartójú eloszlások

1. “Measured but not modeled”
   - dummy modell a lefedetlen részre
2. “Modeled but not measured”

Általános probléma
nemegyenlő tartójú eloszlások összehasonlítása



ábra : nemegyenlő tartójú eloszlások

1. “Measured but not modeled”
   - dummy modell a lefedetlen részre
2. “Modeled but not measured”
   - a KL távolság úgyis bünteti (ki fog jönni)

Bayesi döntéshozatal – I.

- evidence $= \mathbb{P}(\mathcal{H}_i \mid D) = \frac{\mathbb{P}(D|\mathcal{H}_i)\cdot\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(D)}$

Bayesi döntéshozatal – I.

- evidence $= \mathbb{P}(\mathcal{H}_i \mid D) = \frac{\mathbb{P}(D|\mathcal{H}_i)\cdot\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(D)}$
- Ha $\mathbb{P}(\mathcal{H}_i)$ konstans (nincs a priori preferált modell)

$$\mathbb{P}(\mathcal{H}_i \mid D) \propto \mathbb{P}(D \mid \mathcal{H}_i) = \int \mathbb{P}(D \mid \mathbf{w}_i, \mathcal{H}_i) \cdot \mathbb{P}(\mathbf{w}_i \mid \mathcal{H}_i)\, \mathrm{d}\mathbf{w}_i$$

Bayesi döntéshozatal – I.

- evidence $= \mathbb{P}(\mathcal{H}_i \mid D) = \frac{\mathbb{P}(D|\mathcal{H}_i)\cdot\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(D)}$
- Ha $\mathbb{P}(\mathcal{H}_i)$ konstans (nincs a priori preferált modell)

$$\mathbb{P}(\mathcal{H}_i \mid D) \propto \mathbb{P}(D \mid \mathcal{H}_i) = \int \mathbb{P}(D \mid \mathbf{w}_i, \mathcal{H}_i) \cdot \mathbb{P}(\mathbf{w}_i \mid \mathcal{H}_i)\, \mathrm{d}\mathbf{w}_i$$

- Ha $\mathbb{P}(\mathbf{w}_i \mid \mathcal{H}_i)$ konstans (nincs a priori preferált paraméter érték)

$$\frac{1}{\mathrm{Vol}(\mathcal{H}_i)} \cdot \int \mathbb{P}(D \mid \mathbf{w}_i, \mathcal{H}_i)\, \mathrm{d}\mathbf{w}_i$$

Bayesi döntéshozatal – I.

- evidence $= \mathbb{P}(\mathcal{H}_i \mid D) = \frac{\mathbb{P}(D|\mathcal{H}_i)\cdot\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(D)}$
- Ha $\mathbb{P}(\mathcal{H}_i)$ konstans (nincs a priori preferált modell)

$$\mathbb{P}(\mathcal{H}_i \mid D) \propto \mathbb{P}(D \mid \mathcal{H}_i) = \int \mathbb{P}(D \mid \mathbf{w}_i, \mathcal{H}_i) \cdot \mathbb{P}(\mathbf{w}_i \mid \mathcal{H}_i)\, \mathrm{d}\mathbf{w}_i$$

- Ha $\mathbb{P}(\mathbf{w}_i \mid \mathcal{H}_i)$ konstans (nincs a priori preferált paraméter érték)

$$\frac{1}{\mathrm{Vol}(\mathcal{H}_i)} \cdot \int \mathbb{P}(D \mid \mathbf{w}_i, \mathcal{H}_i)\, \mathrm{d}\mathbf{w}_i$$

- Legyen $f(\mathbf{w}_i) = -\frac{1}{n} \ln \mathbb{P}(D \mid \mathbf{w}_i, \mathcal{H}_i)$ az ú.n. *célfüggvény*

Bayesi döntéshozatal – II.

- Laplace integrál közelítő módszer

$$\frac{1}{\mathrm{Vol}(\mathcal{H}_i)} \cdot \mathbb{P}(D \mid \mathbf{w}_i^*, \mathcal{H}_i) \cdot \frac{\left(\frac{2\pi}{n}\right)^{\frac{d}{2}}}{\sqrt{\det f''(\mathbf{w}_i^*)}}$$

Bayesi döntéshozatal – II.

- Laplace integrál közelítő módszer

$$\frac{1}{\mathrm{Vol}(\mathcal{H}_i)} \cdot \mathbb{P}(D \mid \mathbf{w}_i^*, \mathcal{H}_i) \cdot \frac{\left(\frac{2\pi}{n}\right)^{\frac{d}{2}}}{\sqrt{\det f''(\mathbf{w}_i^*)}}$$

- negatív logaritmus $\frac{1}{n}$-szerese:

$$\frac{1}{n} \cdot \ln \mathrm{Vol}(\mathcal{H}_i) + f(\mathbf{w}_i^*) + \frac{1}{2n} \ln \det f''(\mathbf{w}_i^*) + \frac{d}{2n} \cdot \ln \frac{n}{2\pi} \qquad (1)$$

Bayesi döntéshozatal – II.

- Laplace integrál közelítő módszer

$$\frac{1}{\mathrm{Vol}(\mathcal{H}_i)} \cdot \mathbb{P}(D \mid \mathbf{w}_i^*, \mathcal{H}_i) \cdot \frac{\left(\frac{2\pi}{n}\right)^{\frac{d}{2}}}{\sqrt{\det f''(\mathbf{w}_i^*)}}$$

- negatív logaritmus $\frac{1}{n}$-szerese:

$$\frac{1}{n} \cdot \ln \mathrm{Vol}(\mathcal{H}_i) + f(\mathbf{w}_i^*) + \frac{1}{2n} \ln \det f''(\mathbf{w}_i^*) + \frac{d}{2n} \cdot \ln \frac{n}{2\pi} \qquad (1)$$

- kiterjesztett modell

$$\overline{\mathbb{Q}}_{\mathbf{w},q_{d+1},\ldots,q_k}(j) = \begin{cases} \lambda \cdot \mathbb{Q}_{\mathbf{w}}(j) & \text{“measured and modeled”} \\ (1-\lambda) \cdot q_j & \text{“measured but not modeled”} \end{cases}$$

ahol λ a lefedetlen valószínűség

Analízis

- A döntéshozatal függ *n*-től

Analízis

- A döntéshozatal függ *n*-től
- Bünteti a rossz illesztést

Analízis

- A döntéshozatal függ *n*-től
- Bünteti a rossz illesztést
- Bünteti a lefedetlen részt a dummy modell paraméterein keresztül

Analízis

- A döntéshozatal függ *n*-től
- Bünteti a rossz illesztést
- Bünteti a lefedetlen részt a dummy modell paraméterein keresztül
- Bünteti a sok paramétert (paramétertér térfogata és az utolsó tag is)

Analízis

- A döntéshozatal függ n-től
- Bünteti a rossz illesztést
- Bünteti a lefedetlen részt a dummy modell paraméterein keresztül
- Bünteti a sok paramétert (paramétertér térfogata és az utolsó tag is)
- Az egyetlen nehéz tag a $\ln \det f''$

Analízis

- A döntéshozatal függ n-től
- Bünteti a rossz illesztést
- Bünteti a lefedetlen részt a dummy modell paraméterein keresztül
- Bünteti a sok paramétert (paramétertér térfogata és az utolsó tag is)
- Az egyetlen nehéz tag a $\ln\det f''$
- $n \to \infty$ határértékben csak a célfüggvény számít

Analízis

- A döntéshozatal függ n-től
- Bünteti a rossz illesztést
- Bünteti a lefedetlen részt a dummy modell paraméterein keresztül
- Bünteti a sok paramétert (paramétertér térfogata és az utolsó tag is)
- Az egyetlen nehéz tag a $\ln\det f''$
- $n \to \infty$ határértékben csak a célfüggvény számít
- A likelihood:

$$\mathbb{P}(D \mid \mathbf{w}^*, \mathcal{H}_i) = \prod_{j \in \mathrm{supp}(D)} \overline{\mathbb{Q}}_{\mathbf{w}_i^*}(j)^{n_j}$$

Analízis

- A döntéshozatal függ n-től
- Bünteti a rossz illesztést
- Bünteti a lefedetlen részt a dummy modell paraméterein keresztül
- Bünteti a sok paramétert (paramétertér térfogata és az utolsó tag is)
- Az egyetlen nehéz tag a $\ln\det f''$
- $n \to \infty$ határértékben csak a célfüggvény számít
- A likelihood:

$$\mathbb{P}(D \mid \mathbf{w}^*, \mathcal{H}_i) = \prod_{j \in \mathrm{supp}(D)} \overline{\mathbb{Q}}_{\mathbf{w}_i^*}(j)^{n_j}$$

- A célfüggvény a kereszt entrópia

$$-\frac{1}{n} \ln \mathbb{P}(D \mid \mathbf{w}_i^*, \mathcal{H}) = -\sum_{j \in \mathrm{supp}(D)} \frac{n_j}{n} \cdot \ln \overline{\mathbb{Q}}_{\mathbf{w}_i^*}(j)$$

Általánosított KL

- $$-\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j) - \sum_{j \notin} \frac{n_j}{n} \cdot \ln(1-\lambda) \cdot q_j$$

Általánosított KL

- $$-\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j) - \sum_{j \notin} \frac{n_j}{n} \cdot \ln(1-\lambda) \cdot q_j$$
- Elméleti optimum az lenne, ha $\mathbb{Q}_{\mathbf{w}_i^*}(j) = \frac{n_j}{n \cdot \lambda}$ és $q_j = \frac{n_j}{n \cdot (1-\lambda)}$

Általánosított KL

- $$-\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j) - \sum_{j \notin} \frac{n_j}{n} \cdot \ln(1-\lambda) \cdot q_j$$
- Elméleti optimum az lenne, ha $\mathbb{Q}_{\mathbf{w}_i^*}(j) = \frac{n_j}{n \cdot \lambda}$ és $q_j = \frac{n_j}{n \cdot (1-\lambda)}$
- Ekkor az adat entrópiája az optimális érték. Ezért ezt levonhatjuk.

$$\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \frac{\frac{n_j}{n}}{\lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j)} =$$

Általánosított KL

- $$-\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j) - \sum_{j \notin} \frac{n_j}{n} \cdot \ln(1-\lambda) \cdot q_j$$
- Elméleti optimum az lenne, ha $\mathbb{Q}_{\mathbf{w}_i^*}(j) = \frac{n_j}{n \cdot \lambda}$ és $q_j = \frac{n_j}{n \cdot (1-\lambda)}$
- Ekkor az adat entrópiája az optimális érték. Ezért ezt levonhatjuk.

$$\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \frac{\frac{n_j}{n}}{\lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j)} =$$

Általánosított KL

- $$-\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j) - \sum_{j \notin} \frac{n_j}{n} \cdot \ln(1-\lambda) \cdot q_j$$
- Elméleti optimum az lenne, ha $\mathbb{Q}_{\mathbf{w}_i^*}(j) = \frac{n_j}{n \cdot \lambda}$ és $q_j = \frac{n_j}{n \cdot (1-\lambda)}$
- Ekkor az adat entrópiája az optimális érték. Ezért ezt levonhatjuk.

$$\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \frac{\frac{n_j}{n}}{\lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j)} =$$

$$-\lambda \cdot \ln \lambda + \sum_{j \in \{\text{közös tartó}\}} \mathbb{P}(j \mid D) \cdot \ln \frac{\mathbb{P}(j \mid D)}{\mathbb{Q}_{\mathbf{w}_i^*}(j)} \tag{2}$$

- Folytonosra is működik

Általánosított KL

- $$-\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j) - \sum_{j \notin} \frac{n_j}{n} \cdot \ln(1-\lambda) \cdot q_j$$
- Elméleti optimum az lenne, ha $\mathbb{Q}_{\mathbf{w}_i^*}(j) = \frac{n_j}{n \cdot \lambda}$ és $q_j = \frac{n_j}{n \cdot (1-\lambda)}$
- Ekkor az adat entrópiája az optimális érték. Ezért ezt levonhatjuk.

$$\sum_{j \in \{\text{közös tartó}\}} \frac{n_j}{n} \cdot \ln \frac{\frac{n_j}{n}}{\lambda \cdot \mathbb{Q}_{\mathbf{w}_i^*}(j)} =$$

$$-\lambda \cdot \ln \lambda + \sum_{j \in \{\text{közös tartó}\}} \mathbb{P}(j \mid D) \cdot \ln \frac{\mathbb{P}(j \mid D)}{\mathbb{Q}_{\mathbf{w}_i^*}(j)} \tag{2}$$

- Folytonosra is működik
- Egészen addig, amíg

$$\lambda > \frac{1}{e}$$

Eredmények

- ...

[1] Fisher metrika

Eredmények

- ...
- Túl kicsi adatméretnél valószínűleg nem működik a Laplace közelítés

[1] Fisher metrika

Eredmények

- ...
- Túl kicsi adatméretnél valószínűleg nem működik a Laplace közelítés
- Túl nagy adatméretnél nem számít a modellméret

[1] Fisher metrika

Eredmények

- ...
- Túl kicsi adatméretnél valószínűleg nem működik a Laplace közelítés
- Túl nagy adatméretnél nem számít a modellméret
- Az adatméret az egyetlen[1] hiper-paraméter (viszont nem kell újraszámolni semmit, utólag állítható)

[1] Fisher metrika

Eredmények

- ...
- Túl kicsi adatméretnél valószínűleg nem működik a Laplace közelítés
- Túl nagy adatméretnél nem számít a modellméret
- Az adatméret az egyetlen[1] hiper-paraméter (viszont nem kell újraszámolni semmit, utólag állítható)
- 1k-1M adatméret jónak tűnt (?)

[1] Fisher metrika

Eredmények

- ...
- Túl kicsi adatméretnél valószínűleg nem működik a Laplace közelítés
- Túl nagy adatméretnél nem számít a modellméret
- Az adatméret az egyetlen[1] hiper-paraméter (viszont nem kell újraszámolni semmit, utólag állítható)
- 1k-1M adatméret jónak tűnt (?)
- Általában működik bármilyen modellre, akkor is ha az eloszlás tartója modell függő (de egy adott modellre nem lehet paraméterfüggő)

[1] Fisher metrika