BAYESOWSKIE PORÓWNYWANIE MODELI I ŁĄCZENIE WIEDZY

Łukasz Kwiatkowski

Katedra Ekonometrii i Badań Operacyjnych

Plan wykładu

1) Bayesowskie porównywanie modeli – ogólne informacje

2) Przykład – prosty, jednoparametryczny model bayesowski

3) Specyfikacja rozkładu *a priori* na przestrzeni modeli

4) Porównywanie modeli parami

5) Bayesowskie łączenie wiedzy

Porównywanie modeli na gruncie „klasycznej" statystyki/ekonometrii:

→ pod względem dobroci dopasowania do danych (ang. *data fit, in-sample fit*), czy też inaczej – mocy wyjaśniającej /objaśniającej dane zjawisko (ang. *explanatory power*)

→ przeprowadzane za pomocą…?

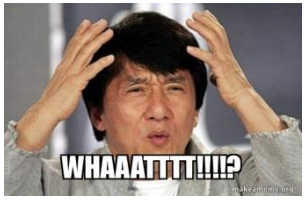
Bayesowskie porównywanie modeli 1

- Przyjmijmy, że do opisu tego samego zjawiska empirycznego, reprezentowanego przez (dokładnie!) ten sam wektor obserwacji y, rozważamy m alternatywnych modeli bayesowkich $M_1, M_2, \dots, M_m$, utożsamianych z łącznymi rozkładami obserwacji, y, oraz parametrów, $\theta^{(i)} \in \Theta^{(i)} \subseteq \mathbb{R}^{\dim \theta^{(i)}}$, $i = 1, 2, \dots, m$:

$$p\left(y, \theta^{(i)} \middle| M_i\right) = p\left(y \middle| \theta^{(i)}, M_i\right) p(\theta^{(i)} | M_i) \qquad (1)$$

(dotychczas nie rozważaliśmy kontekstu wielu modeli, więc nie było potrzeby warunkowania względem M_i)

  - $\theta^{(i)}$ – wektor parametrów i-tego modelu (obejmujący parametry swoiste danego modelu, jak i także – o ile występują – parametry wspólne)
  - $p\left(y \middle| \theta^{(i)}, M_i\right)$ – rozkład próbkowy (funkcja wiarygodności) w modelu M_i
  - $p(\theta^{(i)} | M_i)$ – rozkład *a priori* w modelu M_i
  - M_i – **„etykieta” *i*-tego modelu** – na gruncie bayesowskim jest traktowana jako konkretna (i-ta) wartość **zmiennej losowej** M („M jak model”… ;)



→ Równanie (1) w „porządnym” zapisie powinno wyglądać tak:

$$p\left(y, \theta^{(i)} \middle| M = M_i\right) = p\left(y \middle| \theta^{(i)}, M = M_i\right) p(\theta^{(i)} | M = M_i)$$

Zwykle jednak pomijamy fragment „$M =$”

- Zatem etykieta modelu, M, jest tu zmienną losową (oczywiście, dyskretną) o wartościach w zbiorze $\mathbb{M} = \{M_1, M_2, \dots, M_m\}$

Bayesowskie porównywanie modeli 2

- Zakładamy, że $\mathbb{M} = \{M_1, M_2, \dots, M_m\}$ stanowi **kompletny** zbiór **parami wykluczających się** modeli bayesowskich:
  - **zbiór modeli „parami wykluczających się" (ang. *non-nested models*)**

  $\rightarrow$ **Wykluczamy** zatem **modele zagnieżdżone**, tzn. żaden M_i ($i \in \{1, \dots, m\}$) nie może być szczególnym przypadkiem któregoś spośród pozostałych (**ALE UWAGA**: z dokładnością do zbiorów miary zero w rozkładzie *a priori*) – przykład: $M_1: y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$ *vs.* $M_2: y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$

  $\rightarrow$ M_1 JEST zagnieżdżony w M_2 (poprzez oczywistą restrykcję: $\phi_2 = 0$), ale jeśli $p(\phi_2|M_2)$ jest rozkładem absolutnie ciągłym, wówczas $\Pr(\phi_2 = 0|M_2) = 0$ (= pole pod $p(\phi_2|M_2)$ w argumencie $\phi_2 = 0$), więc z probabilistycznego punktu widzenia, te modele możemy traktować jako niezagnieżdżone

  - **„kompletny" zbiór modeli (ang. *exhaustive set of models)***

  $\rightarrow$ Skład zbioru $\mathbb{M}$, owszem, możemy zmienić (np. uwzględnić jakąś dodatkową specyfikację), ale konieczne będzie przeprowadzenie rachunków „od nowa", uwzględniających „nowy" skład rozważanego zbioru modeli

Bayesowskie porównywanie modeli 3

➢ Modele $M_1, M_2, \dots, M_m$ mogą się między sobą różnić na DWA sposoby:

- Z uwagi na <u>rozkład *a priori*</u>:

  Przykład: Rozważmy dwa modele bayesowskie dla próby $y_t|\lambda \sim iiExp(\lambda)$ $(t = 1, 2, \dots, T)$ z różnymi rozkładami *a priori* (oba jednak – dla uproszczenia – z tej samej rodziny rozkładów gamma – sprzężony z rozważaną funkcją wiarygodności)

  → $p(y|M_1) = p(y|M_2) = \lambda^T e^{-\lambda \sum_{t=1}^{T} y_t}$ (ten sam rozkład próbkowy…)

  → $p(\lambda|M_1) = f_G(\lambda|a_1, b_1), p(\lambda|M_2) = f_G(\lambda|a_2, b_2)$ (…ale różne rozkłady *a priori*)

  (λ – parametr wspólny w obydwu modelach, więc nie wymaga indeksu i)

- Z uwagi na <u>rozkład próbkowy:</u>

  Przykład: Ciąg dziennych logarytmicznych stóp zwrotu z jakiegoś instrumentu finansowego, $y_t, t = 1, 2, \dots, T$:

  M_1: $y_t|(\mu, \sigma^2) \sim iiN(\mu, \sigma^2) \quad \Rightarrow \theta^{(1)} = (\mu, \sigma^2)'$

  M_2: $y_t|(\nu, m, \tau) \sim iit(m, \tau, \nu) \Rightarrow \theta^{(2)} = (\nu, m, \tau)'$

  →m – parametr niecentralności (modalna), τ – precyzja, ν – liczba stopni swobody

  (różne rozkłady próbkowe – z różnymi parametrami, więc i rozkłady *a priori* będą inne; choć po części mogą się pokrywać, np. te dla μ i ν)

Bayesowskie porównywanie modeli 4

- Skoro „etykieta” modelu, M, jest zmienną losową (o wartościach w $\mathbb{M} = \{M_1, M_2, \dots, M_m\}$), to możemy mówić o:
  - $\Pr(M = M_i) \equiv \Pr(M_i)$ – p-stwo *a priori* modelu M_i
  - $\Pr(M = M_i|y) \equiv \Pr(M_i|y)$ – p-stwo *a posteriori* modelu M_i

- Bayesowskie porównywanie (mocy wyjaśniającej) modeli – za pomocą ich p-stw *a posteriori*, $\Pr(M_i|y) = \cdots$ (o którym za chwilę)

- Prawdopodobieństwa *a priori* i – osobno – *a posteriori* tworzą dyskretne rozkłady prawdopodobieństwa zmiennej losowej M na przestrzeni modeli, $\mathbb{M}$:
  - $\Pr(M_i) > 0 \ \wedge \ \sum_{i=1}^{m} \Pr(M_i) = 1$ ($\Pr(M_i) = 0$ tylko dla modeli spoza $\mathbb{M}$)
  - $\Pr(M_i|y) > 0 \ \wedge \ \sum_{i=1}^{m} \Pr(M_i|y) = 1$

Bayesowskie porównywanie modeli 5

➢ Prawdopodobieństwo *a posteriori* modelu M_i:

$$\Pr(M_i|y) = \frac{p(y, M_i)}{p(y)} = \frac{p(y|M_i)\Pr(M_i)}{\sum_{i=1}^{m} p(y|M_i)\Pr(M_i)}$$

- $\Pr(M_i)$ – określamy samodzielnie na całej przestrzeni modeli $\mathbb{M}$, tak aby zachodziła koniunkcja:
  - dla każdego $i = 1, 2, \dots, m$ zachodziło $\Pr(M_i) > 0$ (żaden z modeli w $\mathbb{M}$ nie jest wykluczony *a priori*)
  - $\sum_{i=1}^{m} \Pr(M_i) = 1$ ($\mathbb{M}$ stanowi kompletną klasę rozważanych modeli bayesowskich)

- $p(y|M_i)$ – wartość brzegowej gęstości wektora obserwacji w i-tym modeli = „dotychczasowe $p(y)$" (dotychczas nie rozważaliśmy kontekstu wielu modeli, więc nie było potrzeby warunkowania względem M_i):

$$p(y|M_i) = \int_{\Theta^{(i)}} p\left(y, \theta^{(i)}|M_i\right) d\theta^{(i)} = \int_{\Theta^{(i)}} p\left(y|\theta^{(i)}, M_i\right) p\left(\theta^{(i)}|M_i\right) d\theta^{(i)}$$

lub też – przekształcając wzór Bayesa:

$$p\left(\theta^{(i)}|y, M_i\right) = \frac{p\left(y, \theta^{(i)}|M_i\right)}{p(y|M_i)} \propto p\left(y, \theta^{(i)}|M_i\right)$$

otrzymujemy:

$$p(y|M_i) = \frac{p\left(y, \theta^{(i)}|M_i\right)}{p\left(\theta^{(i)}|y, M_i\right)}$$

→ Wyższa wartość $p(y|M_i)$ dla konkretnego y oznacza wyższe „szanse" tego, że dane „zostały wygenerowane" przez model M_i („pochodzą" z modelu M_i) => dane bardziej wspierają model M_i

(ang. *the data provide more evidence for model* M_i)

- $p(y)$ – wartość brzegowej gęstości wektora obserwacji „uśredniona po modelach" (czyli po wycałkowaniu – tu raczej „wysumowaniu" – niepewności związanej z wyborem „prawidłowej" specyfikacji modelowej):

$$p(y) = \sum_{i=1}^{m} p(y, M_i) = \sum_{i=1}^{m} p(y|M_i)\Pr(M_i)$$

Bayesowskie porównywanie modeli 6

➢ **Trzy uwagi:**

1) Terminologiczna:

- $p(y|M_i)$, $p(y)$ – wartość funkcji gęstości brzegowego rozkładu wektora obserwacji; w skrócie: brzegowa gęstość obserwacji
- Terminy anglojęzyczne:
  - *marginal data density*, MDD;
  - nieco mniej poprawnie: *marginal likelihood* → *likelihood* = wiarygodność, a ta w statystyce jest definiowana jako funkcja *parametrów*, a nie danych; z drugiej strony

$$p(y|M_i) = \underbrace{\int_{\Theta^{(i)}} \underbrace{p\left(y|\theta^{(i)}, M_i\right)}_{r.\ próbkowy} \underbrace{p\left(\theta^{(i)}|M_i\right)}_{r.\ a\ priori} d\theta^{(i)}}_{aka\ f.\ wiarygodności} \quad \text{"ubrzegawianie" } f.\ wiarygodności$$

2) Związek rozkładu *a priori* z wartością MDD:

→ Im rozkład *a priori* jest bliższy funkcji wiarygodności (informacja *a priori* o parametrach jest bliższa tej, którą niosą ze sobą dane), tym wartość MDD jest wyższa ⇒ „Zarzut”: Poprzez „odpowiednie” dobranie rozkładu *a priori* mogę „podbić” dopasowanie modelu → kwestia (nie)uczciwości badawczej…

3) MDD daje się wyznaczyć analitycznie tylko w prostych modelach; zwykle jej wartość trzeba aproksymować/szacować metodami numerycznymi (niełatwymi!), o których – przy innej okazji

Bayesowskie porównywanie modeli 7

➢ **Podsumowując** – chcąc przeprowadzić porównanie mocy wyjaśniającej różnych, alternatywnych modeli bayesowskich:

1) W każdym modelu potrzebujemy wyznaczyć wartość **brzegowej gęstości obserwacji**:

$$p(y|M_i) = \int_{\Theta^{(i)}} p\left(y, \theta^{(i)}|M_i\right) d\theta^{(i)}$$

lub

$$p(y|M_i) = \frac{p\left(y, \theta^{(i)}|M_i\right)}{p(\theta^{(i)}|y, M_i)} = \frac{p\left(y|\theta^{(i)}, M_i\right) p\left(\theta^{(i)}|M_i\right)}{p(\theta^{(i)}|y, M_i)}$$

2) Każdemu z modeli potrzebujemy przypisać **p-stwo *a priori*, $\Pr(M_i)$**, $i = 1, \dots, m$

→ o różnych podejściach – później

→ jeśli nie zakłada się inaczej, wówczas zwykle przyjmuje się równe wartości: $\Pr(M_i) = \frac{1}{m}$ (czyli rozkład jednostajny na przestrzeni modeli)

3) Wyznaczamy **p-stwa *a posteriori*** poszczególnych modeli:

$$\Pr(M_i|y) = \frac{p(y, M_i)}{p(y)} = \frac{p(y|M_i)\Pr(M_i)}{\sum_{i=1}^{m} p(y|M_i)\Pr(M_i)}$$

→ na ich podstawie możemy skonstruować **ranking** modeli (pod względem ich dobroci dopasowania do danych)

Przykład: prosty, jednoparametryczny model bayesowski

- **Kluczowy poblem:** wyznaczenie $p(y|M_i)$ (tj. wartości brzegowej gęstości obserwacji w *i*-tym modelu)
- **Przećwiczmy to** w jakimś prostym, jednoparametrycznym modelu:

→ Na razie bez kontekstu wielu modeli

- $y_t|\lambda \sim iid\, Exp(\lambda),\ \lambda>0 \Rightarrow p(y|\lambda) = \lambda^T e^{-\lambda \sum_{t=1}^{T} y_t}$
- $p(\lambda) = f_G(\lambda|a,b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$

$$p(y,\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{T+a-1} e^{-\lambda\left(b+\sum_{t=1}^{T} y_t\right)}$$

→ $p(\lambda|y) = f_G(\lambda|\bar{a},\bar{b}) = \frac{\bar{b}^{\bar{a}}}{\Gamma(\bar{a})} \lambda^{\bar{a}-1} e^{-\bar{b}\lambda}$, $\bar{a} = T+a$, $\bar{b} = b + \sum_{t=1}^{T} y_t$

$$\Rightarrow p(y) = \frac{p(y,\lambda)}{p(\lambda|y)} = \frac{b^a}{\Gamma(a)} \cdot \frac{\Gamma(\bar{a})}{\bar{b}^{\bar{a}}} = \frac{b^a}{\bar{b}^{\bar{a}}} \cdot \frac{\Gamma(\bar{a})}{\Gamma(a)}$$

→ Gdybyśmy chcieli porównać dwa modele - różniące się tylko wartościami apriori:

$$\lambda|M_1 \sim G(a_1,b_1),\ \lambda|M_2 \sim G(a_2,b_2) \Rightarrow p(\lambda|y,M_i) = f_G(\lambda|\bar{a}_i,\bar{b}_i),\ i=1,2$$

→ Wtedy: $p(y|M_i) = \frac{b_i^{a_i}}{\bar{b}_i^{\bar{a}_i}} \cdot \frac{\Gamma(\bar{a}_i)}{\Gamma(a_i)}$, $i=1,2$

$$\bar{a}_i = T + a_i$$
$$\bar{b}_i = b_i + \sum_{t=1}^{T} y_t$$

Specyfikacja rozkładu *a priori* na przestrzeni modeli 1

- Prawdopodobieństwa *a priori* porównywanych modeli muszą spełniać **3 postulaty**:

1) $\Pr(M_i) > 0, i = 1, 2, \dots, m$

→ wszystkie rozważane modele się „liczą” *a priori*, żadnego z nich *a priori* nie wykluczamy

2) $\sum_{i=1}^{m} \Pr(M_i) = 1$

→ … a zarazem „liczą” się tylko(!) te modele, i żadne inne spoza klasy $\mathbb{M} = \{M_1, M_2, \dots, M_m\}$

3) $\forall_{\substack{i,j \in \{1,2,\dots,m\} \\ i \neq j}} \Pr(M_i \vee M_j) = \Pr(M_i) + \Pr(M_j)$

→ modele są parami niezagnieżdżone

Specyfikacja rozkładu *a priori* na przestrzeni modeli 2

- Dwie „szkoły” ustalania prawdopodobieństw *a priori* modeli $M_1, M_2, \dots, M_m$:

  1) Po prostu, przyjmujemy równe ich wartości (rozkład jednostajny na przestrzeni modeli):

$$\Pr(M_i) = \frac{1}{m}, \qquad i = 1, 2, \dots, m$$

  Zauważmy, że wtedy:

$$\Pr(M_i|y) = \frac{p(y, M_i)}{p(y)} = \frac{p(y|M_i)\Pr(M_i)}{\sum_{j=1}^{m} p(y|M_j)\Pr(M_j)} = \frac{p(y|M_i)}{\sum_{j=1}^{m} p(y|M_j)}$$

Specyfikacja rozkładu *a priori* na przestrzeni modeli 3

- Dwie szkoły ustalania prawdopodobieństw *a priori* modeli $M_1, M_2, \ldots, M_m$:

  2) **Premiujemy** *a priori* modele oszczędnie sparametryzowane (o **mniejszej** liczbie parametrów) – jeśli jakieś dwa modele jednakowo wyjaśniają dane zjawisko (mają tę samą wartość $p(y|M_i)$), wówczas *a priori* faworyzujemy ten, który jest prostszy (osiąga to samo dopasowanie mniejszym „kosztem")

  $\rightarrow$ William Ockham – *brzytwa Ockhama* (ang. *Ockham's razor*):

  „*Nie należy mnożyć bytów ponad potrzebę*"

  $\rightarrow$ Arnold Zellner – *reguła KISS* (ang. *Keep It Sophistically Simple*)

  ALE: Nie oznacza to, że z góry odrzucamy modele bardziej skomplikowane – owszem, bierzemy je pod uwagę, ale *a priori* przypisujemy im relatywnie mniejsze szanse => dane muszą nas „mocno przekonać", że takie bardziej skomplikowane, rozbudowane specyfikacje „opłacają się", są „niezbędne".

  Czasami także, poprzez przyjęcie odpowiednio wyższego p-stwa *a priori*, uwzględniamy sytuację, w której pewien model posiada w jakimś sensie „uprzywilejowany status teoretyczny" (np. jest przedstawicielem jakiejś ogólnie przyjętej teorii, „wszyscy w niego wierzą"). Pozostałą masę p-stwa rozdzielamy pomiędzy pozostałymi modelami.

Specyfikacja rozkładu *a priori* na przestrzeni modeli 4

- Przykładowa reguła „premiowania” modeli prostszych:

  Niech $l_i = \dim \Theta^{(i)} =$ liczba parametrów modelu M_i. Prawdopodobieństwa ustalamy wg formuły:

$$\Pr(M_i) = \frac{2^{-l_i}}{\sum_{j=1}^{m} 2^{-l_j}} \propto 2^{-l_i}$$

  $\rightarrow$ Im większe l_i, tym mniejsze $\Pr(M_i)$

- **Przykład**: Załóżmy, że rozważane są trzy modele bayesowskie, M_1, M_2, M_3, o liczbie parametrów, odpowiednio: $l_1 = 3, l_2 = 5, l_3 = 4$

  Obliczamy wartości wyrażenia 2^{-l_i} i dzielimy je przez ich sumę:

i	1	2	3	Suma w wierszu
l_i	3	5	4	
2^{-l_i}	…	…	…	…
$\Pr(M_i)$	…	…	…	= 1

  $\rightarrow$ *Again*: Im większe l_i, tym mniejsze $\Pr(M_i)$

Porównywanie modeli parami 1

- Chcemy porównać **relatywną** moc wyjaśniającą **dwóch modeli**: $M_i, M_j \in \mathbb{M} = \{M_1, \dots, M_m\}$
- Iloraz szans *a posteriori* (ang. *posterior odds ratio*, POR):

$$POR_{ij} = \frac{\Pr(M_i|y)}{\Pr(M_j|y)} = \frac{\frac{\Pr(M_i)\,p(y|M_i)}{p(y)}}{\frac{\Pr(M_j)\,p(y|M_j)}{p(y)}} = \underbrace{\frac{\Pr(M_i)}{\Pr(M_j)}}_{\substack{iloraz\ szans\ a\ priori\ (ang.\ prior\ odds\ ratio)}} \cdot \underbrace{\frac{p(y|M_i)}{p(y|M_j)}}_{\substack{czynnik\ Bayesa\ (ang.\ Bayes\ factor)}}$$

$\rightarrow$ Informuje o tym, ilokrotnie wyższe/niższe jest p-stwo *a posteriori* modelu M_i (w liczniku) w stosunku do modelu M_j (w mianowniku)

$\rightarrow POR_{ij} \in (0, +\infty)$:

- $0 < POR_{ij} < 1$: Model M_i jest mniej prawdopodobny *a posteriori* od modelu M_j
- $POR_{ij} = 1$: Oba modele są jednakowo prawdopodobne *a posteriori*
- $POR_{ij} > 1$: Model M_i jest bardziej prawdopodobny *a posteriori* od modelu M_j

Porównywanie modeli parami 2

- Z poprzedniej strony: Iloraz szans *a posteriori* (ang. *posteriori odds ratio*, POR):

$$POR_{ij} = \frac{\Pr(M_i|y)}{\Pr(M_j|y)} = \frac{\frac{\Pr(M_i)\,p(y|M_i)}{p(y)}}{\frac{\Pr(M_j)\,p(y|M_j)}{p(y)}} = \underbrace{\frac{\Pr(M_i)}{\Pr(M_j)}}_{\substack{iloraz\ szans \\ a\ priori}} \cdot \underbrace{\frac{p(y|M_i)}{p(y|M_j)}}_{\substack{czynnik \\ Bayesa}}$$

- Iloraz szans *a priori* (ang. *prior odds ratio*, PROR):

$$PROR_{ij} = \frac{\Pr(M_i)}{\Pr(M_j)}$$

→ „Ilokrotnie bardziej/mniej wierzę *a priori* w model M_i – w stosunku do modelu M_j”

- Czynnik Bayesa (ang. *Bayes factor*, BF):

$$BF_{ij} = \frac{p(y|M_i)}{p(y|M_j)}$$

→ Informuje o tym, ilokrotnie wyższą/niższą szansę pojawienia się zaobserwowanego wektora y dawał model M_i w stosunku do modelu M_j

→ $BF_{ij} \in (0, +\infty)$: analogicznie jak w przypadku POR_{ij}

- Zwykle zakłada się rozkład jednostajny na przestrzeni modeli ($\forall_{i,j=1,\ldots,m} \Pr(M_i) = \Pr(M_j)$):

$$POR_{ij} = BF_{ij} \quad \rightarrow \quad POR \text{ i } BF \text{ informują o tym samym}$$

Porównywanie modeli parami 3

- POR, BF – własności:

1) Jeżeli $\Pr(M_i) = \Pr(M_j)$, wówczas $POR_{ij} = BF_{ij}$

2) $POR_{ij} = \frac{1}{POR_{ji}}, \qquad BF_{ij} = \frac{1}{BF_{ji}}$

3) Dla dowolnego $l \in \{1, 2, \dots, m\}$: $\quad BF_{ij} = \frac{BF_{il}}{BF_{jl}} = \frac{BF_{li}}{BF_{lj}}$

4) Do obliczenia $\Pr(M_i|y)$ wystarczy znajomość (wszystkich) czynników Bayesa BF_{jl} $(j = 1, 2, \dots, m)$ dla danego, dowolnie wybranego $l \in \{1, 2, \dots, m\}$:

$$\Pr(M_i|y) = \frac{p(y|M_i)\Pr(M_i)}{\sum_{j=1}^{m} p(y|M_j)\Pr(M_j)} = \frac{p(y|M_i)\Pr(M_i)}{\sum_{j=1}^{m} p(y|M_j)\Pr(M_j)} \cdot \frac{\frac{1}{p(y|M_l)}}{\frac{1}{p(y|M_l)}} = \frac{\Pr(M_i)\frac{p(y|M_i)}{p(y|M_l)}}{\sum_{j=1}^{m} \Pr(M_j)\frac{p(y|M_j)}{p(y|M_l)}}$$

$$= \frac{\Pr(M_i)\, BF_{il}}{\sum_{j=1}^{m} \Pr(M_j) BF_{jl}}$$

→ $BF_{jj} = 1$

5) Przy założeniu $\forall_{i=1,\dots,m}\ \Pr(M_i) = \frac{1}{m}$:

$$\Pr(M_i|y) = \frac{BF_{il}}{\sum_{j=1}^{m} BF_{jl}}$$

Porównywanie modeli parami 4

- **Ćwiczenie 1**: Wyobraźmy sobie, że porównujemy pewne dwa modele bayesowskie: M_1 i M_2. Zinterpretuj poniższe wartości:
  - $POR_{12} = 2{,}5$ $\rightarrow$ Model M_1 jest 2,5 razy (czyli o 150%) bardziej prawdopodobny *a posteriori* od modelu M_2.
  - $POR_{12} = 0{,}25$ $\rightarrow$ Łatwiej tu zinterpretować $POR_{\mathbf{21}} = \frac{1}{POR_{12}} = 4$: Model $M_{\mathbf{2}}$ jest 4-krotnie (czyli o 300%) bardziej prawdopodobny *a posteriori* od modelu $M_{\mathbf{1}}$.
  - $POR_{12} = 1{,}33$ $\rightarrow$ Model M_1 jest o 33% bardziej prawdopodobny *a posteriori* od modelu M_2 (lepiej to brzmi, aniżeli stwierdzenie, że „Model M_1 jest 1,33 razy bardziej prawdopodobny…”)

Porównywanie modeli parami 5

- **Logarytmy <u>dziesiętne</u> POR, BF**: $\log POR_{ij} \equiv \log_{10} POR_{ij}, \qquad \log BF_{ij} \equiv \log_{10} BF_{ij}$

  $\rightarrow$ Informują o różnicy w rzędzie wielkości

  $\rightarrow \log POR_{ij} \in \mathbb{R}$:

  - $\log POR_{ij} < 0 \Leftrightarrow POR_{ij} < 1$: Model M_i jest mniej prawdopodobny *a posteriori* od modelu M_j
  - $\log POR_{ij} = 0 \Leftrightarrow POR_{ij} = 1$: Oba modele są jednakowo prawdopodobne *a posteriori*
  - $\log POR_{ij} > 0 \Leftrightarrow POR_{ij} > 1$: Model M_i jest bardziej prawdopodobny *a posteriori* od modelu M_j

  $\rightarrow \log POR_{ij} = -\log POR_{ji}$

- **Przykład:** Przyjmijmy, że rozważamy trzy modele: M_1, M_2, M_3 o prawdopodobieństwach *a posteriori*, odpowiednio: $Pr(M_1|y) = 0{,}6794;$ $\Pr(M_2|y) = 0{,}0006;$ $\Pr(M_3|y) = 0{,}32$

  - $POR_{12} \approx 1132{,}333 = 1{,}132333 \cdot 10^3 \Rightarrow \log POR_{12} \approx 3{,}054 > 0$ $\rightarrow$ Interpretacja…
  - $POR_{13} \approx 2{,}123 = 2{,}123 \cdot 10^0 \Rightarrow \log POR_{13} \approx 0{,}327 > 0$ $\rightarrow$ Interpretacja…
  - $POR_{23} \approx 0{,}00188 = 1{,}88 \cdot 10^{-3} \Rightarrow \log POR_{23} \approx -2{,}727 \approx -3 < 0$ $\rightarrow$ Interpretacja…

Bayesowskie łączenie wiedzy 1 – NIE OBOWIĄZUJE (ale jest ciekawe ;)

- **Dwa konteksty rozważania grupy modeli (a nie tylko jednej specyfikacji):**

  1) Gdy testujemy konkurujące między sobą modele (lub teorie, które leżą u ich podstaw)

     → Celem rozważania różnych modeli jest wówczas wskazanie/wybór tego jednego – najlepszego

     → Najlepszy model – ten charakteryzujący się najwyższą wartością p-stwa *a posteriori* (lub najwyższą wartością *MDD* – przy równych p-stwach *a priori* wszystkich rozważanych modeli)

  2) Gdy chcemy wnioskować bądź to o pewnych **wspólnych** dla tych modeli parametrach czy wielkościach (będących funkcjami tych pierwszych), bądź dokonać prognozy z ich wykorzystaniem

     → Zamiast wybierać, a tym samym ograniczać się tylko do jednego (najlepszego) modelu, można „połączyć wiedzę" ze wszystkich modeli

     ⇒ ***Bayesowskie łączenie wiedzy:*** połączenie uzyskanych w poszczególnych modelach indywidualnych rozkładów *a posteriori* danej wielkości w jeden, „ostateczny" rozkład *a posteriori*; w celu uzyskania tego rozkładu należy „wycałkować" niepewność związaną z wyborem „prawidłowego" modelu

- **Terminy anglojęzyczne**:
  - *Bayesian pooling approach (BPA)*
  - *Bayesian model averaging (BMA)* ← nie do końca poprawne

Bayesowskie łączenie wiedzy 2

- Niech $\lambda \in \Lambda \subseteq \mathbb{R}^{\dim \Lambda}$ oznacza wektor pewnych wspólnych wielkości w modelach $M_1, M_2, \dots, M_m$, o rozkładach *a posteriori* $p(\lambda|y, M_i)$, $i = 1, 2, \dots, m$

  $\rightarrow$ Oczywiście, w szczególnym przypadku λ może być także skalarem

  $\rightarrow$ λ – może reprezentować:

  - bezpośrednio, wspólne *parametry* grupy modeli

    Przykład: $M_1: y_t = \beta_0 + \beta_1 x_{t1} + \varepsilon_t$; $M_2: y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \varepsilon_t$

    → $\lambda = (\beta_0 \ \beta_1 \ \sigma^2)'$ bądź $\lambda = \beta_1$

  - jakieś wielkości, które mają jednakowe (wspólne) znaczenie/interpretację w poszczególnych modelach (choć same są różnymi funkcjami parametrów tych modeli)

    Przykład: elastyczność produkcji względem kapitału w modelu Cobba-Douglasa i w modelu Translog ($El_{Q|K}$ ma różną postać funkcyjną w obydwu tych modelach, ale znaczenie/interpretację – to samo) → Zatem tutaj $\lambda = El_{Q|K}$

  - prognozowaną wartość zjawiska (o prognozowaniu bayesowskim – za jakiś czas :)

    Przykład: Modelujemy inflację, y_t, w kolejnych kwartałach $t = 1, 2, \dots, T$. Modelowane dane: $y = (y_1 \ y_2 \ \dots \ y_T)'$. Interesuje nas wnioskowanie o y_{T+1}, czyli właśnie prognozowanie – na gruncie bayesowskim odbywa się to za pomocą tzw. *rozkładu predyktywnego*: $p(y_{T+1}|y)$, czyli rozkładu *a posteriori* obserwacji niedostępnej, prognozowanej. Taki rozkład możemy zbudować w różnych modelach zbudowanych dla y: $p(y_{T+1}|y, M_i)$, $i = 1, \dots, m$

    → Zatem tutaj $\lambda = y_{T+1}$

Bayesowskie łączenie wiedzy 3

- Od strony operacyjnej, celem bayesowskiego łączenia wiedzy o λ jest wyznaczenie jednego, „ostatecznego”, „końcowego” rozkładu *a posteriori* analizowanej wielkości – bezwarunkowego względem modelu, $p(\lambda|y)$ – na podstawie indywidualnych, uzyskanych w poszczególnych modelach rozkładów *a posteriori*, $p(\lambda|y, M_i)$, $i = 1, 2, \dots, m$:

$$p(\lambda|y) = \sum_{i=1}^{m} p(\lambda, M_i|y) = \sum_{i=1}^{m} p(\lambda|y, M_i) \Pr(M_i|y)$$

→ „Wycałkowujemy” zmienną losową oznaczającą model

→ $p(\lambda|y)$ jest dyskretną mieszanką (ang. *mixture*) rozkładów $p(\lambda|y, M_i)$, $i = 1, 2, \dots, m$, przy czym rozkładem „mieszającym” (wagami) jest tu rozkład *a posteriori* na przestrzeni modeli

→ Bayesowskie łączenie wiedzy o λ polega na „uśrednieniu” indywidualnych gęstości *a posteriori* $p(\lambda|y, M_i)$, $i = 1, 2, \dots, m$, z wagami równymi prawdopodobieństwom *a posteriori* poszczególnych modeli, $\Pr(M_i)$, $i = 1, 2, \dots, m$.

→ $p(\lambda|y)$ uwzględnia w sobie niepewność związaną z wyborem „prawidłowej” specyfikacji modelowej

→ BPA „opłaca się” w szczególności wtedy, gdy p-stwa *a posteriori* modeli są dość zbliżone

Bayesowskie łączenie wiedzy 4

➢ Wzór z poprzedniej strony:

$$p(\lambda|y) = \sum_{i=1}^{m} p(\lambda, M_i|y) = \sum_{i=1}^{m} p(\lambda|y, M_i)\,\Pr(M_i|y)$$

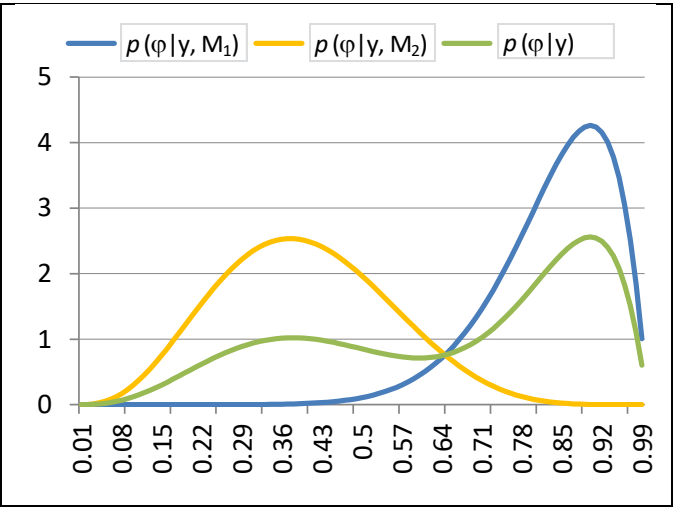
→ To gotowy przepis na formułę funkcji gęstości tego „ostatecznego” rozkładu *a posteriori* dla λ

Przykład: Wyobraźmy sobie, że mamy dwa modele, M_1 i M_2, zbudowane dla prostej próby losowej z rozkładu dwupunktowego, $y_t|\varphi \sim iiBern(\varphi)$ [model prób Bernoulliego]. W obydwu modelach przyjęto sprzężony rozkład *a priori* (beta, o „jakichś tam” hiperparametrach) i uzyskano rozkłady *a posteriori*: $p(\varphi|y, M_1) = f_{Be}(\varphi|\overline{a_1} = 10, \overline{b_1} = 2)$ oraz $p(\varphi|y, M_2) = f_{Be}(\varphi|\overline{a_2} = 4, \overline{b_2} = 6)$. Wyznaczono także prawdopodobieństwa *a posteriori* obydwu modeli: $\Pr(M_1|y) = 0{,}6$; $\Pr(M_2|y) = 1 - \Pr(M_1|y) = 0{,}4$. Jak wygląda „ostateczny” rozkład *a posteriori* parametru φ?

$$p(\varphi|y) = \Pr(M_1|y)\,p(\varphi|y, M_1) + \Pr(M_2|y)\,p(\varphi|y, M_2)$$

$$= 0{,}6 \cdot f_{Be}(\varphi|\overline{a_1} = \cdots, \overline{b_1} = \cdots) + 0{,}4 \cdot f_{Be}(\varphi|\overline{a_2} = \cdots, \overline{b_2} = \cdots)$$

$$= 0{,}6 \cdot \frac{1}{B(\overline{a_1}, \overline{b_1})}\varphi^{\overline{a_1}-1}(1-\varphi)^{\overline{b_1}-1} + 0{,}4 \cdot \frac{1}{B(\overline{a_2}, \overline{b_2})}\varphi^{\overline{a_2}-1}(1-\varphi)^{\overline{b_2}-1}$$



Bayesowskie łączenie wiedzy 5

➢ **Charakterystyki „końcowego” rozkładu *a posteriori***

- **Funkcja gęstości** – patrz wyżej
- **Wartość oczekiwana**

$$E(\lambda|y) = \int_{\Lambda} \lambda p(\lambda) d\lambda = \int_{\Lambda} (\lambda \sum_{i=1}^{m} p(\lambda|y, M_i) \Pr(M_i|y)) d\lambda =$$

$$\sum_{i=1}^{m} \left(\underbrace{\Pr(M_i|y)}_{\substack{nie\ zależy \\ od\ \lambda}} \underbrace{\int_{\Lambda} \lambda p(\lambda|y, M_i) d\lambda}_{E(\lambda|y, M_i)} \right) = \sum_{i=1}^{m} E(\lambda|y, M_i) \Pr(M_i|y)$$

→ $E(\lambda|y)$ jest średnią ważoną indywidualnych, otrzymanych w poszczególnych modelach wartości oczekiwanych, z wagami równymi p-stwom *a posteriori* modeli

- **Wariancja**

$$Var(\lambda|y) = \overbrace{E\left[\left(\lambda - E(\lambda|y)\right)^2 \middle| y\right] = E(\lambda^2|y) - E^2(\lambda|y)}^{wzory\ ogólne}$$

$$\underbrace{=}_{\substack{łatwo \\ pokazać}} \sum_{i=1}^{m} \left((Var(\lambda|y, M_i) + E^2(\lambda|y, M_i)) \right) \Pr(M_i|y) - E^2(\lambda|y)$$

- **Pozostałe** – nie da się uzyskać ogólnych formuł ⇒ konieczna numeryczna aproksymacja (oszacowanie) ich wartości na podstawie **próby losowej** wygenerowanej z „końcowego” rozkładu *a posteriori* – patrz kolejna strona

Bayesowskie łączenie wiedzy 6

- **Ogólnie:** numeryczna aproksymacja (oszacowanie) dowolnych charakterystyk dowolnego rozkładu prawdopodobieństwa, dajmy na to $p(x)$ (czyli pewnego rozkładu zmiennej losowej x) – dwa kroki:
  - Generujemy N-elementową próbę (pseudo-)losową z analizowanego rozkładu: $\{x^{(1)}, x^{(2)}, \dots, x^{(N)}\} \equiv \{x^{(q)}\}_{q=1}^{N}$, gdzie $x^{(q)}$ oznacza q-te losowanie (ang. *draw*)
  - Dowolną charakterystykę rozkładu $p(x)$ możemy aproksymować obliczając jej próbkowy odpowiednik (estymator – z daszkiem!) na podstawie uzyskanej próby $\{x^{(q)}\}_{q=1}^{N}$:
    - Funkcję gęstości, $p(x)$, możesz oszacować/przybliżyć za pomocą histogramu (tak, histogram jest estymatorem! :)
    - $E(x) = \int_{\mathbb{R}} xp(x)dx \approx \hat{E}(x) = \frac{1}{N}\sum_{q=1}^{N} x^{(q)} \equiv \bar{x}$
    - $Var(x) = E\big(x - E(x)\big)^2 = \int_{\mathbb{R}}\big(x - E(x)\big)^2 p(x)dx \approx \widehat{Var}(x) = \frac{1}{N}\sum_{q=1}^{N}\big(x^{(q)} - \bar{x}\big)^2$

      → Alternatywnie: $\widetilde{Var}(x) = \frac{1}{N-1}\sum_{q=1}^{N}\big(x^{(q)} - \bar{x}\big)^2$, lecz asymptotycznie ($T \to \infty$) oba estymatory wariancji są równoważne
    - Kwantyle – jak na statystyce opisowej → gotowe funkcje w pakietach → *Zadanko*: Sprawdź, czy da się obliczyć w Excelu obliczyć kwantyl (nie kwaRtyl!) dowolnego rzędu, $\alpha \in (0, 1)$
    - Modalna – jak na statystyce opisowej → gotowe funkcje w pakietach → *Zadanko*: Sprawdź, czy da się obliczyć modalną w Excelu (weź pod uwagę, że – w ogólności – zmienna losowa x może być ciągła albo dyskretna!)

Bayesowskie łączenie wiedzy 7

- **<u>W kontekście łączenia wiedzy</u>** – schemat jak powyżej, ale pojawia się…

  → Pytanie: Jak uzyskać próbę losową z „końcowego” rozkładu *a posteriori*, będącego *mieszanką* indywidualnych rozkładów *a posteriori*:

  $$p(\lambda|y) = \sum_{i=1}^{m} p(\lambda, M_i|y) = \sum_{i=1}^{m} p(\lambda|y, M_i)\,\Pr(M_i|y)$$

  Ad 1) Generujemy N-elementową próbę (pseudo-)losową $\{\lambda^{(q)}\}_{q=1}^{N}$ z rozkładu $p(\lambda|y)$ – w tym celu będzie potrzebne wcześniejsze wygenerowanie prób z indywidualnych rozkładów $p(\lambda|y, M_i)$

  1.a) Ustal liczebność próby N z „końcowego” rozkładu *a posteriori* (generalnie: im większa, tym wyższej dokładności będą szacunki)

  1.b) Niech N_i $(i = 1, 2, \dots, m)$ oznacza liczbę losowań dla λ, generowanych z indywidualnych rozkładów *a posteriori* w poszczególnych modelach, $p(\lambda|y, M_i)$. Próby te oznaczmy symbolem $\{\lambda^{(q)}|M_i\}_{q=1}^{N_i}$. Wartości N_i ustalmy tak, aby:

  - $\sum_{i=1}^{m} N_i = N$
  - $\forall_{i=1,2,\dots,m} \frac{N_i}{N} = \Pr(M_i|y)$ → Udziały N_i w N odzwierciedlają p-stwa *a posteriori* modeli
    → Czyli po prostu $N_i = N \cdot \Pr(M_i|y)$

  1.c) Próba z „końcowego” rozkładu *a posteriori*, $\{\lambda^{(q)}\}_{q=1}^{N}$, powstaje poprzez połączenie tych z rozkładów indywidualnych, tj. $\{\lambda^{(q)}\}_{q=1}^{N} = \bigcup_{i=1}^{m} \{\lambda^{(q)}|M_i\}_{q=1}^{N_i}$