*Wnioskowanie bayesowskie w ekonomii empirycznej (*Analityka gospodarcza, zima 2023/2024)

***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*)

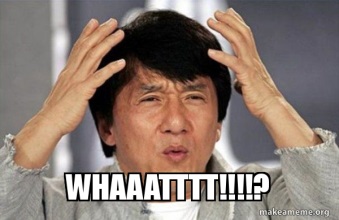
🡪 przeprowadzane za pomocą…?

**Bayesowskie porównywanie modeli 1**

* Przyjmijmy, że do opisu tego samego zjawiska empirycznego, reprezentowanego przez (dokładnie!) ten sam wektor obserwacji , rozważamy alternatywnych modeli bayesowkich , utożsamianych z łącznymi rozkładami obserwacji, , oraz parametrów, , :

(1)

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

* – wektor parametrów -tego modelu (obejmujący parametry swoiste danego modelu, jak i także – o ile występują – parametry wspólne)
* – rozkład próbkowy (funkcja wiarygodności) w modelu
* – rozkład *a priori* w modelu
*  – **„etykieta” -tego modelu** – na gruncie bayesowskim jest traktowana jako konkretna (-ta) wartość **zmiennej losowej** („ jak model”… ;)

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

Zwykle jednak pomijamy fragment „”

* Zatem etykieta modelu, , jest tu zmienną losową (oczywiście, dyskretną) o wartościach w zbiorze

**Bayesowskie porównywanie modeli 2**

* Zakładamy, że stanowi **kompletny** zbiór **parami wykluczających się** modeli bayesowskich:
  + **zbiór modeli „parami wykluczających się” (ang. *non-nested models*)**

→ **Wykluczamy** zatem **modele zagnieżdżone**, tzn. żaden () nie może być szczególnym przypadkiem któregoś spośród pozostałych (**ALE UWAGA**: z dokładnością do zbiorów miary zero w rozkładzie *a priori*) – przykład: : *vs.* :

🡪 JEST zagnieżdżony w (poprzez oczywistą restrykcję: ), ale jeśli jest rozkładem absolutnie ciągłym, wówczas (= pole pod w argumencie ), więc z probabilistycznego punktu widzenia, te modele możemy traktować jako niezagnieżdżone

* + **„kompletny” zbiór modeli (ang. *exhaustive set of models)***

→ Skład zbioru , 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 mogą się między sobą różnić na DWA sposoby:
  + Z uwagi na rozkład *a priori*:

**Przykład**: Rozważmy dwa modele bayesowskie dla próby () z różnymi rozkładami *a priori* (oba jednak – dla uproszczenia – z tej samej rodziny rozkładów gamma – sprzężony z rozważaną funkcją wiarygodności)

🡪 (ten sam rozkład próbkowy…)

🡪 , (…ale różne rozkłady *a priori*)

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

* + Z uwagi na rozkład próbkowy:

**Przykład**: Ciąg dziennych logarytmicznych stóp zwrotu z jakiegoś instrumentu finansowego, :

⇒

⇒

🡪 – parametr niecentralności (modalna), – precyzja, – 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 i )

**Bayesowskie porównywanie modeli 4**

* Skoro „etykieta” modelu, , jest zmienną losową (o wartościach w ), to możemy mówić o:
  + – p-stwo *a priori* modelu
  + – p-stwo *a posteriori* modelu
* Bayesowskie porównywanie (mocy wyjaśniającej) modeli – za pomocą ich p-stw *a posteriori*, (o którym za chwilę)
* Prawdopodobieństwa *a priori* i – osobno – *a posteriori* tworzą dyskretne rozkłady prawdopodobieństwa zmiennej losowej na przestrzeni modeli, :
  + ( tylko dla modeli spoza )

**Bayesowskie porównywanie modeli 5**

* Prawdopodobieństwo *a posteriori* modelu :
  + – określamy samodzielnie na całej przestrzeni modeli , tak aby zachodziła koniunkcja:
    - dla każdego zachodziło (żaden z modeli w nie jest wykluczony *a priori*)
    - ( stanowi kompletną klasę rozważanych modeli bayesowskich)
  + – wartość brzegowej gęstości wektora obserwacji w -tym modeli = „dotychczasowe ” (dotychczas nie rozważaliśmy kontekstu wielu modeli, więc nie było potrzeby warunkowania względem ):

lub też – przekształcając wzór Bayesa:

otrzymujemy:

🡪 Wyższa wartość dla konkretnego oznacza wyższe „szanse” tego, że dane „zostały wygenerowane” przez model („pochodzą” z modelu ) => dane bardziej wspierają model

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

* + – wartość brzegowej gęstości wektora obserwacji „uśredniona po modelach” (czyli po wycałkowaniu – tu raczej „wysumowaniu” – niepewności związanej z wyborem „prawidłowej” specyfikacji modelowej):

.

**Bayesowskie porównywanie modeli 6**

* **Trzy uwagi:**

1. Terminologiczna:
   * , – 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
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…

1. 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**:

lub

1. Każdemu z modeli potrzebujemy przypisać **p-stwo *a priori*,** ,

🡪 o różnych podejściach – później

🡪 jeśli nie zakłada się inaczej, wówczas zwykle przyjmuje się równe wartości:

(czyli rozkład jednostajny na przestrzeni modeli)

1. Wyznaczamy **p-stwa *a posteriori***poszczególnych modeli:

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

**Przykład: prosty, jednoparametryczny model bayesowski**

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



**Specyfikacja rozkładu *a priori* na przestrzeni modeli 1**

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

1. ,

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



→ … a zarazem „liczą” się tylko(!) te modele, i żadne inne spoza klasy



→ modele są parami niezagnieżdżone

**Specyfikacja rozkładu *a priori* na przestrzeni modeli 2**

* Dwie „szkoły” ustalania prawdopodobieństw *a priori* modeli :

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

,

Zauważmy, że wtedy:

**Specyfikacja rozkładu *a priori* na przestrzeni modeli 3**

* Dwie szkoły ustalania prawdopodobieństw *a priori* modeli :

1. **Premiujemy** *a priori* modele oszczędnie sparametryzowane (o **mniejszej** liczbie parametrów) – jeśli jakieś dwa modele jednakowo wyjaśniają dane zjawisko (mają tę samą wartość ), wówczas *a priori* faworyzujemy ten, który jest prostszy (osiąga to samo dopasowanie mniejszym „kosztem”)

→ William Ockham – *brzytwa Ockhama* (ang. *Ockham’s razor*):

„*Nie należy mnożyć bytów ponad potrzebę*”

→ 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 liczba parametrów modelu . Prawdopodobieństwa ustalamy wg formuły:

→ Im większe , tym mniejsze

* **Przykład**: Załóżmy, że rozważane są trzy modele bayesowskie, , o liczbie parametrów, odpowiednio: , ,

Obliczamy wartości wyrażenia i dzielimy je przez ich sumę:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | Suma w wierszu |
|  | 3 | 5 | 4 |
|  | … | … | … | … |
|  | … | … | … | = 1 |

→ *Again*: Im większe , tym mniejsze

**Porównywanie modeli parami 1**

* Chcemy porównać **relatywną** moc wyjaśniającą **dwóch modeli**:
* Iloraz szans *a posteriori* (ang. *posterior odds ratio*, POR):

→ Informuje o tym, ilokrotnie wyższe/niższe jest p-stwo *a posteriori* modelu (w liczniku) w stosunku do modelu (w mianowniku)

→ :

* Model jest mniej prawdopodobny *a posteriori* od modelu
* Oba modele są jednakowo prawdopodobne *a posteriori*
* Model jest bardziej prawdopodobny *a posteriori* od modelu

**Porównywanie modeli parami 2**

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

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

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

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

→ : analogicznie jak w przypadku

* Zwykle zakłada się rozkład jednostajny na przestrzeni modeli ():

→ i informują o tym samym

**Porównywanie modeli parami 3**

* – własności:

1. Jeżeli , wówczas
2. ,
3. Dla dowolnego :
4. Do obliczenia wystarczy znajomość (wszystkich) czynników Bayesa () dla danego, dowolnie wybranego :

🡪

1. Przy założeniu :

**Porównywanie modeli parami 4**

* **Ćwiczenie 1**: Wyobraźmy sobie, że porównujemy pewne dwa modele bayesowskie: i . Zinterpretuj poniższe wartości:
  + → Model M1 jest 2,5 razy (czyli o 150%) bardziej prawdopodobny *a posteriori* od modelu M2.
  + → Łatwiej tu zinterpretować : Model M**2** jest 4-krotnie (czyli o 300%) bardziej prawdopodobny *a posteriori* od modelu M**1**.
  + → Model M1 jest o 33% bardziej prawdopodobny *a posteriori* od modelu M2 (lepiej to brzmi, aniżeli stwierdzenie, że „Model M1 jest 1,33 razy bardziej prawdopodobny…”)

**Porównywanie modeli parami 5**

* **Logarytmy dziesiętne POR, BF**: ,

→ Informują o różnicy w rzędzie wielkości

→ :

* + : Model jest mniej prawdopodobny *a posteriori* od modelu
  + : Oba modele są jednakowo prawdopodobne *a posteriori*
  + : Model jest bardziej prawdopodobny *a posteriori* od modelu

→

* **Przykład:** Przyjmijmy, że rozważamy trzy modele: o prawdopodobieństwach *a posteriori*, odpowiednio: ; ;
  + ⇒ → Interpretacja…
  + ⇒ → Interpretacja…
  + ⇒ → 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)

1. 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 oznacza wektor pewnych wspólnych wielkości w modelach , o rozkładach *a posteriori* ,

→ Oczywiście, w szczególnym przypadku może być także skalarem

→ – może reprezentować:

* bezpośrednio, wspólne *parametry* grupymodeli

Przykład: ;

🡪 bądź

* 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 ( ma różną postać funkcyjną w obydwu tych modelach, ale znaczenie/interpretację – to samo) 🡪 Zatem tutaj

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

Przykład: Modelujemy inflację, , w kolejnych kwartałach . Modelowane dane: . Interesuje nas wnioskowanie o , czyli właśnie prognozowanie – na gruncie bayesowskim odbywa się to za pomocą tzw. *rozkładu predyktywnego*: , czyli rozkładu *a posteriori* obserwacji niedostępnej, prognozowanej. Taki rozkład możemy zbudować w różnych modelach zbudowanych dla : ,

🡪 Zatem tutaj

**Bayesowskie łączenie wiedzy 3**

* Od strony operacyjnej, celem bayesowskiego łączenia wiedzy o jest wyznaczenie jednego, „ostatecznego”, „końcowego” rozkładu *a posteriori* analizowanej wielkości – bezwarunkowego względem modelu, – na podstawie indywidualnych, uzyskanych w poszczególnych modelach rozkładów *a posteriori*, , :

🡪 „Wycałkowujemy” zmienną losową oznaczającą model

🡪 jest dyskretną mieszanką (ang. *mixture*) rozkładów , , przy czym rozkładem „mieszającym” (wagami) jest tu rozkład *a posteriori* na przestrzeni modeli

🡪 Bayesowskie łączenie wiedzy o polega na „uśrednieniu” indywidualnych gęstości *a posteriori* , , z wagami równymi prawdopodobieństwom *a posteriori* poszczególnych modeli, , .

🡪 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:

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

Przykład: Wyobraźmy sobie, że mamy dwa modele, i , zbudowane dla prostej próby losowej z rozkładu dwupunktowego, [model prób Bernoulliego]. W obydwu modelach przyjęto sprzężony rozkład *a priori* (beta, o „jakichś tam” hiperparametrach) i uzyskano rozkłady *a posteriori*: oraz . Wyznaczono także prawdopodobieństwa *a posteriori* obydwu modeli: ; . Jak wygląda „ostateczny” rozkład *a posteriori* parametru ?

|  |  |
| --- | --- |
|  |  |

**Bayesowskie łączenie wiedzy 5**

* **Charakterystyki „końcowego” rozkładu *a posteriori***
  + **Funkcja gęstości** – patrz wyżej
  + **Wartość oczekiwana**

🡪 jest średnią ważoną indywidualnych, otrzymanych w poszczególnych modelach wartości oczekiwanych, z wagami równymi p-stwom *a posteriori* modeli

* + **Wariancja**

* + **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 (czyli pewnego rozkładu zmiennej losowej ) – dwa kroki:
  + Generujemy -elementową próbę (pseudo-)losową z analizowanego rozkładu: , gdzie oznacza -te losowanie (ang. *draw*)
  + Dowolną charakterystykę rozkładu możemy aproksymować obliczając jej próbkowy odpowiednik (estymator – z daszkiem!) na podstawie uzyskanej próby :
    - Funkcję gęstości, , możesz oszacować/przybliżyć za pomocą histogramu (tak, histogram jest estymatorem! :)

🡪 Alternatywnie: , lecz asymptotycznie () 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,
    - 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 może być ciągła albo dyskretna!)

**Bayesowskie łączenie wiedzy 7**

* **W kontekście łączenia wiedzy** – schemat jak powyżej, ale pojawia się…

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

Ad 1) Generujemy -elementową próbę (pseudo-)losową z rozkładu – w tym celu będzie potrzebne wcześniejsze wygenerowanie prób z indywidualnych rozkładów

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

1.b) Niech () oznacza liczbę losowań dla , generowanych z indywidualnych rozkładów *a posteriori* w poszczególnych modelach, . Próby te oznaczmy symbolem . Wartości ustalmy tak, aby:

* → Udziały w odzwierciedlają p-stwa *a posteriori* modeli

🡪 Czyli po prostu

1.c) Próba z „końcowego” rozkładu *a posteriori*, , powstaje poprzez połączenie tych z rozkładów indywidualnych, tj.