Model Chandrasekhara / Smoluchowskiego - 1 pudełko

Piotr Piękos

22 sierpnia 2019

1 Oznaczenia

W celach notacyjnych rozbijemy proces $X(t)$ (ilość żyjących osób) na dwa procesy:

- $N(t)$ - Ilość narodzin
- $S(t)$ - Ilość śmierci

Wtedy $X(t) = N(t) - S(t)$, Dodatkowo oznaczymy intensywności procesów przez:

- a_N - intensywność procesu narodzin
- a_S - parametr rozkładu wykładniczego odpowiadającego za długość życia

Dodatkowe oznaczenia:

- $I_X(t)$ - indeksy "żywych" zmiennych w momencie t.
- W_i - zmienna losowa (o rozkładzie wykladniczym z parametrem a_S) mówiąca o długości życia osoby i

Możnaby spróbować zamodelować $S(t)$ jako niejednorodny proces Poissona z intensywnością zależną od $N(t)$. Ja jednak to rozdzieliłem jedynie ze względów notacyjnych.

2 Prawa ewolucji

$P(X(t+h) = x+1|X(t) = x)$:

Korzystamy tutaj z faktu, że dla procesu Poissona (N) mamy:

- $P(N(t+h) = n+1|N(t) = n) = a_N h + o(h)$
- $P(N(t+h) \geqslant n+2|N(t) = n) = o(h)$

Dodatkowo:

- $P(N(t+h) = n|N(t) = n) = 1 - a_N h + o(h)$
- $P(S(t+h) = s|S(t) = s, X(t) = x) = P(\forall_i \in I_X(t) W_i \geqslant h) + o(h) = \prod_{i \in I_X(t)} P(W_i \geqslant h) + o(h) = e^{-a_S x h} + o(h) = 1 - a_S x h + o(h)$
- $P(S(t+h) = s+1|S(t) = s, X(t) = x) = x(e^{-a_S(x-1)h} - e^{-a_S x h}) + o(h) = a_S x h + o(h)$
- $P(S(t+h) = s+2|S(t) = s, X(t) = x) = o(h)$

$o(h)$ pojawia się już po pierwszej równości ze względu na to, że przy dokładnym rozpisaniu prawdopodobienstw należałoby warunkować w którym momencie X(t) się zmieni (X jest zależny od S), jednak ta różnica jest $o(h)$, więc po prostu jest zawarta w tym.

zatem

$$P(X(t+h) = x+1|X(t) = x) =$$
$$P(N(t+h) = n+1|N(t) = n) \cdot P(S(t+h) = s|S(t) = s, X(t) = x) =$$
$$(a_N h + o(h)) \cdot (1 - a_S x h + o(h)) =$$
$$a_N h + o(h)$$

$$P(X(t+h) = x-1|X(t) = x) =$$
$$P(N(t+h) = n|N(t) = n) \cdot P(S(t+h) = s+1|S(t) = s, X(t) = x) =$$
$$(1 - a_N h + o(h)) \cdot (a_S x h + o(h)) =$$
$$a_S x h + o(h)$$

Czyli mamy

- $P(X(t+h) = x+1 | X(t) = x) = a_N h + o(h)$
- $P(X(t+h) = x-1 | X(t) = x) = a_S x h + o(h)$

wzory te razem z $Q(x,x) = -a_N h - a_S x h$ i zerami w pozostałych wierszach opisują intensywność przejść procesu Markowa

3 Symulacje

3.1 Algorytm

Algorytm symulacji składa się z dwóch części:

1. standardowa symulacja procesu Poissona (czasy narodzin)
2. symulacja czasów życia z rozkładu wykładniczego o parametrze a_S

Konkretnie:

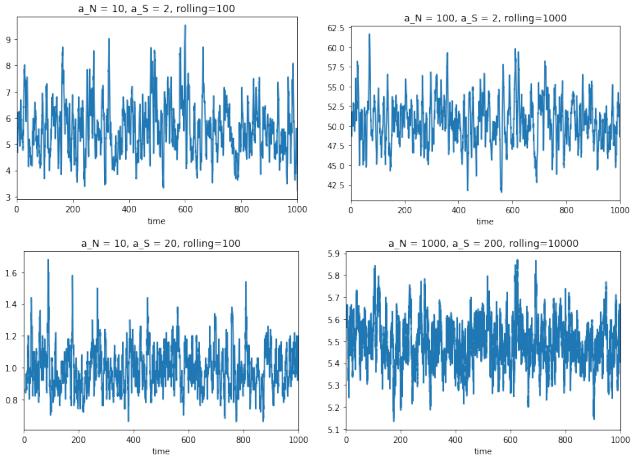
Gen $N \sim Poiss(a_N t)$
for $i = 1$ to N do Gen $U_i \sim U(0,t)$, Gen $L_i \sim Exp(1/a_S)$
$(T_1, ..., T_n) = \text{Sort}(U_1, ..., U_n)$ otrzymujemy proces Poissona, czasy narodzin
$(D_1, ..., D_n) = (T_1 + L_1, ..., T_n + L_n)$ - dodajemy niezalezne czasy zycia do czasow narodzin i mamy czasy smierci.

Algorytm korzysta z gotowych bibliotek (numpy) do symulacji rozkładów Poissona i rozkładu wykładniczego. Dodatkowo wykorzystuje w nich możliwość wektoryzacji. Ale można to zrobić surowo od rozkładu jednostajnego np. za pomocą metody odwracania dystrybuanty oraz chociażby metody eliminacji.

3.2 Przykładowe trajektorie

Wygenerowane wykresy są sparametryzowane 4 parametrami:

- t - długość symulacji



Rysunek 1: Przykładowe trajektorie symulacji

- $a\_N - a_N$
- $a\_S - a_S$
- rolling - długość horyzontu średniej kroczącej, średnia krocząca jest konieczna dla wielu wykresów ze względu na ogromną ilośc punktów na wykresie. Pokazuje jednak ona "gęstość" punktów.

W trajektoriach należy także zwrócić uwagę na skalę, gdyż ona odgrywa kluczową rolę.

Od razu widać, że wyróżnioną liczbą jest $\frac{a_N}{a_S}$. Proces oscyluje w jej okolicach, co jest zrozumiałe po spojrzeniu na prawa ewolucji. Mówią one, że punkt $\frac{a_N}{a_S}$ jest punktem granicznym dla którego intensywność śmierci jest taka sama jak intensywność narodzin.

Możnaby pokusić się o alternatywną parametryzację procesu za pomocą parametrów $a = \frac{a_N}{a_S}$ i $b = a_S$, wtedy a reprezentowałaby punkt w okół którego symulacja będzie oscylować, a b reprezentowałby gęstość punktow.

Skyupmy się teraz na dwóch procesach o tym samym parametrze $a(\frac{a_N}{a_S}$. Są widoczne na pierwszym i ostatnim wykresie wyżej ($\frac{a_N}{a_S} = 5$). Na wykresach widać po pierwsze większy parametr rolling który jest konieczny ze względu na większa gęstość punktów. Z tych wykresów możnaby też odczytać, że gęstszy wykres ma mniejsze odchylenia, jednak byłoby to błędne, gdyż jest to spowodowane jedynie wygładzeniem przez średnią kroczącą.

4 Analiza

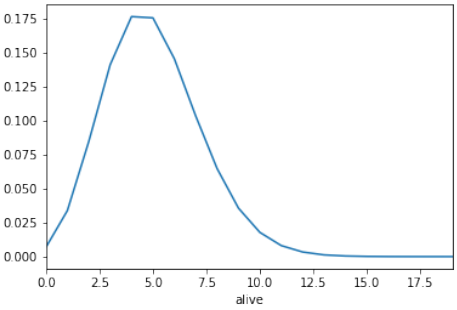
4.1 Rozkładu stacjonarnego

Empiryczną masę rozkładu można łatwo zdefiniować przez częstość występowania procesu w danym stanie (gdzie stan jest zdefiniowany przez ilość żywych osób).

czyli

$$P(X = n) = \frac{\int \mathbb{I}_{\{X(t)=n\}} dt}{t}$$

Symulacyjnie rozkład jest stabilny od parametrów, dodatkowo statystyką dostateczną dla rozkładu jest $\frac{a_N}{a_S}$.



Rysunek 2: Przykładowa symulacja dla $a_N = 10, a_S = 2, t = 100000$

4.2 Wartość oczekiwana

Mając empiryczny rozkład prawdopodobieństwa wartość oczekiwana jest prostolinijna do policzenia za pomocą tradycyjnego wzoru na wartość oczekiwaną.

z symulacji jasno wynika, że

$$\lim_{t\to\infty} \mathbb{E}X = \frac{a_N}{a_S}$$

4.3 Postać rozkładu stacjonarnego i testy statystyczne

Rozkład stacjonarny można wyprowadzić z zależności $\pi Q = 0$ lub zgadnąć na podstawie symulacji.

$$\pi = Poiss(\frac{a_N}{a_S})$$

Aby sprawdzić ten rezultat wykonałem test Kołmogorova-Smirnova. Porównałem w nim rozkład empiryczny (podany wyżej) z oczekiwanym rozkładem - $Poiss(\frac{a_N}{a_S})$

Interesuje nas istotność statystyczna na poziomie $\alpha = 0.05$. Z tabelek testu KS wnioskujemy, że wartość graniczna dla nas to $\frac{1.36}{\sqrt{n}}$.

Żeby sprawdzić hipotezę potrzebne jest policzenie statystyki Kołmogorowa: $\sup_n |F_n(x) - F(x)|$. Standardowo w przypadku gdy jest ona większa niż wartość graniczna odrzucamy tezę, a w przypadku gdy jest mniejsza wnioskujemy, że nie ma istotnych statystycznie różnic i nie można wnioskować, że te rozkłady są różne. Uznamy to za akceptację naszej tezy.

Do celów testów został zasymulowany proces przez 100000 jednostek czasu z parametrami $a_N = 10, a_S = 5$. Sprawdzamy więc, czy rozkład jest statystycznie różny od $Poiss(2)$

$$\sup_n |F_n(x) - F(x)| = 0.000793 < 0.000961 = \frac{1.36}{\sqrt{n}}$$

Zakładam tutaj, że n to ilość przejść procesu - w tym przypadku 2001615. Nie jestem jednak przekonany co do słuszności tego testu, jako, że takie n nie uwzględnia czasów trwania w poszczególnych stanach.

todo: sprawdzenie innych $a_N i a_S$

Skoro test nie pokazał istotnej różnicy między rozkładami uznajemy, że nasz rozkład stacjonarny jest postaci $Poiss(2)$ dla $a_N = 10, a_S = 2$

5 Uruchamianie symulacji

kod do uruchamiania symulacji - w folderze jedno_pudla::

```
$ python3 sym.py t a_N a_S output_file
```

gdzie:

- t - czas trwania symulacji
- a_N - intensywność procesu narodzin
- a_S - Parametr rozkładu czasu życia
- output_file - nazwa pliku do którego ma się zapisać symulacja (csv)

wykresy zostały stworzone w notebooku (jedno_pudlo/)simulate.ipynb