Analiza bayesowska, modelowanie danych

Plan wykładu

- analiza Bayesowska
- modelowanie danych doświadczalnych
- łączenie danych, metoda podstawowa i metoda Kalmana

literatura:

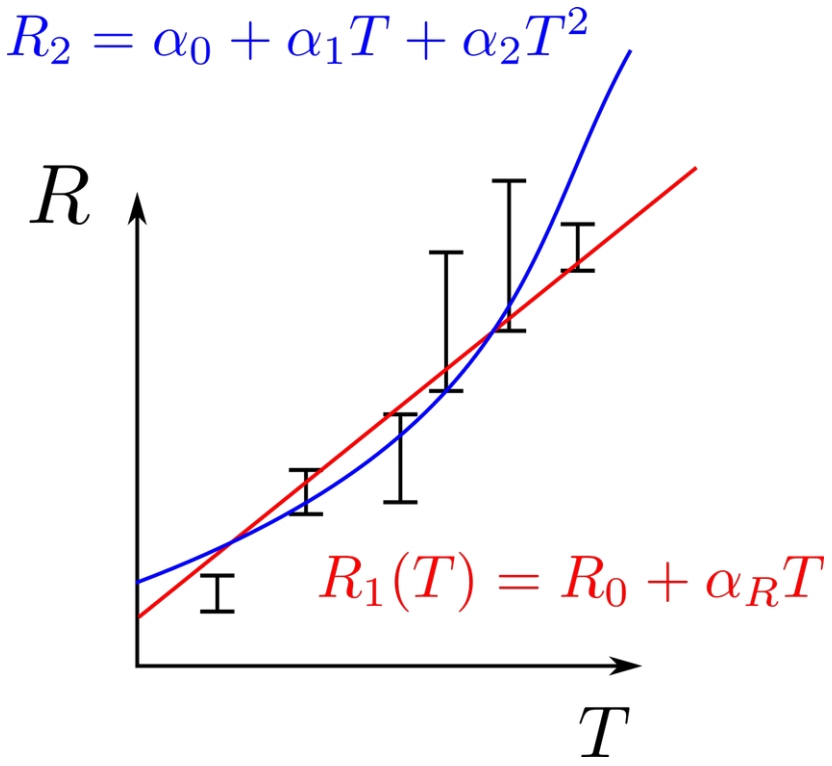
- J.F. Boudreau, E.S. Swanson, „Applied Computational Physics”
- D.S. Sivia, J. Skilling, „Data analysis. A Bayesian tutorial”

Eksperyment wykonujemy w celu zebrania danych, a same dane musimy przetworzyć do postaci, która będzie zawierać tylko istotne dla nas informacje.

Prosty przykład z pracowni fizycznej: **mierzymy opór próbki w funkcji temperatury**

- jeśli próbka to konwencjonalny przewodnik to wiemy, że opór opisuje równanie

$$R(T) = R_0 + \alpha_R \cdot T$$



i jedynie musimy określić współczynniki tego równania

$$R_0 \pm \sigma_{R_0}, \qquad \alpha_R \pm \sigma_{\alpha_R}$$

- w pewnych przypadkach nie potrafimy określić wszystkich procesów wpływających na wynik i w związku z tym nie dysponujemy odpowiednim modelem, w takim przypadku zazwyczaj proponujemy pewien model parametryczny i staramy się określić jego parametry,

  problem polega na tym, że alternatywnych modeli możemy zaproponować wiele, zatem bazując na danych doświadczalnych powinniśmy określić jakość naszego modelu w języku rachunku prawdopodobieństwa/statystyki

Analiza Bayesowska

Najprostsza intuicyjnie analiza prawdopodobieństwa występowania zdarzeń losowych w praktyce polega na obserwacji danego typu zdarzenia, zarejestrowaniu pewnej liczby przypadków i na podstawie zaobserwowanej liczby interesujących zdarzeń określamy częstość ich występowania, np. losujemy N razy kostką i sprawdzamy jak często wypada cyfra 5

$$\eta_5 = \frac{n_5}{N} \stackrel{?}{\approx} P\{5\}$$

Bazujemy tu na pojedynczej informacji, co może bardzo ograniczyć wiarygodność wyniku.

Przykład: wykonujemy N=10 rzutów, 5 wypada tylko raz n_5=1, jaki wniosek?

$$P\{5\} \neq \frac{1}{10} = 0.2$$

Oczywiście to czego dokonaliśmy jest nadużyciem, ponieważ tylko w granicy N → ∞ możemy postawić znak równości, dla mało-licznych zbiorów danych określanie prawdopodobieństwa na podstawie częstości ich występowania może prowadzić do błędnych wniosków. Aby tego uniknąć uzyskać bardziej wiarygodny wynik musimy do naszego stanu wiedzy o danym zdarzeniu dodać dodatkowe informacje – które analiza częstościowa zupełnie ignoruje – tym zajmuje się analiza bayesowska.

Interesuje nas zdarzenie losowe A, jak je opisujemy?

P{A} – prawdopodobieństwo wystąpienia zdarzenia A, więcej nie wiemy

P{A|O} – pr. warunkowe wystąpienia A, gdy zdarzenie O wystąpi wcześniej
tu inwestujemy w większy zasób informacji początkowych

P{A+B|O} – pr. warunkowe wystąpienia A lub B, gdy wystąpiło O

P{AB|O} - pr. wystąpienia A oraz B, gdy wystąpiło O

$P\{\bar{A}|O\} + P\{A|O\} = 1$ - zdarzenie A jest realizowane z pewnym pr. ale jeśli nie zajdzie to realizowane jest zdarzenie przeciwne (brak A), oba zdarzenia wykluczają się wzajemnie i tworzą zupełną przestrzeń zdarzeń

$P\{A\bar{A}\} = 0$ - nie da się rzucić monetą {orzeł, reszka} i dostać jednocześnie dwóch wyników, realizowane jest tylko jedno zdarzenie

Prawdopodobieństwo wystąpienia dwóch zdarzeń możemy określić z reguły iloczynowej

$$P\{AB|O\} = P\{A|BO\}P\{B|O\}$$

- czyli najpierw określamy pr. wystąpienia B (rozkład marginalny/zredukowany), następnie określamy pr. wystąpienia A pod warunkiem, że wystąpiło B (zdarzenie pewne)

$$\int P\{AB|O\}dA = P\{B|O\}$$

- rozkład zredukowany dostajemy po wysumowaniu/całkowaniu wkładów od wszystkich możliwych realizacji A

Ponieważ, kolejność wystąpienia A i B nie jest ściśle określona czynnikami zewnętrznymi więc alternatywnie możemy zapisać odwracając kolejność zdarzeń

$$P\{BA|O\} = P\{B|AO\}P\{A|O\}$$

porównujemy ze sobą oba wyrażenia

$$P\{AB|O\} = P\{BA|O\}$$

$$P\{A|BO\}P\{B|O\} = P\{B|AO\}P\{A|O\}$$

dostajemy wzór będący sednem twierdzenia Bayesa

$$P\{A|BO\} = P\{A|O\}\frac{P\{B|AO\}}{P\{B|O\}}$$

zapiszmy wzór w bardziej praktycznej postaci

Przyjmujemy oznaczenia

$$A \equiv \vec{\theta}$$

- wektor nieznanych wartości parametrów dla danego modelu danych

$$B \equiv \vec{D} = \{x_1, x_2, \ldots, x_N\}$$

- wektor danych pomiarowych

$$P\{\vec{\theta}|DO\} = P\{\vec{\theta}|O\}\frac{P\{D|\vec{\theta}O\}}{P\{D|O\}}$$

$P\{\theta|\vec{D}O\}$

- „posterior probability”, określamy pr. że szukany (i znaleziony) zestaw parametrów uzyskamy dla zestawu danych D

$P\{\theta|\vec{O}\}$

- „prior probability”, to nasza informacja (niepełna, przybliżona, w zasadzie nieznana) o rozkładzie parametrów θ

$P\{D|\theta\vec{O}\}$

- „likelihood function”, **funkcja wiarygodności**, to ona pozwoli nam znaleźć wektor parametrów θ, tu określa prawdopodobieństwo, że dla danego zestawu parametrów θ uda nam się odtworzyć wektor danych pomiarowych D

$P\{D|O\}$

- pr. zredukowane, otrzymania zestawu danych pomiarowych D, w praktyce pełni ono rolę czynnika normalizacyjnego

Nasza dotychczasowa praktyka podpowiada nam że zamiast operować pojęciem prawdopodobieństwa=dystrybuantą, wygodniej jest posługiwać się fgp. Pomijając jawną zależność od informacji o eksperymencie (zdarzenie O – eksperyment został wykonany), analogiczne wyrażenie dla fgp ma postać

$$f(\vec{\theta}|D) = g(\vec{\theta})\frac{L(D|\vec{\theta})}{M}$$

$$M = \int_{\Theta} L(D|\vec{\theta})g(\vec{\theta})d^m\theta$$

Czyli aby określić rozkład parametrów θ tak aby były one zgodne z danymi D musimy podać:

- rozkład g(θ)
- funkcję wiarygodności L(D|**θ**)
- znajomość czynnika normalizacyjnego M (po lewej i po prawej stronie równania mamy fgp unormowane) zazwyczaj nie jest potrzebna

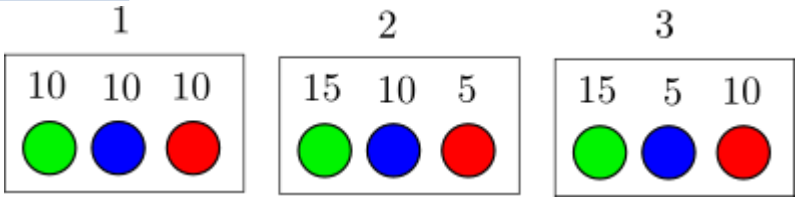
Uwagi:

- dane D i wektor parametrów θ są traktowane jako dowolnie liczne zmienne losowe, o ciągłym lub dyskretnym rozkładzie
- fgp dla parametrów to f(), dla danych fgp jest proporcjonalne do L() (M w mianowniku normalizuje iloczyn, a nie L)
- rozkładu g(θ) zazwyczaj nie znamy nawet w przybliżeniu, więc opisujemy go jakąś rozsądną funkcją ale to oznacza, że musi dopasować się funkcja wiarygodności, tak aby równanie było spełnione
- sposób definiowania funkcji wiarygodności determinuje też sposób określania optymalnego wektora θ co stanowi sedno analizy bayesowskiej

Przykład - określanie prawdopodobieństwa w modelu tradycyjnym i bayesowskim.

Mamy 3 urny, w każdej jest po 30 kolorowych kul (rysunek).
Z tego układu losujemy jedną kulę.

Wylosowaliśmy zieloną kulę, ale nie wiemy z której urny.
Jakie jest prawdopodobieństwo że kula pochodzi z urny 1,2, lub 3?



tradycyjne podejście: podzielić liczbę kul zielonych (nzi) w urnie przez całkowitą liczbę kul występujących w 3 urnach

$$P\{1|z\} = \frac{10}{40} = 0.25 \qquad P\{2|z\} = \frac{15}{40} = 0.375 \qquad P\{3|z\} = \frac{15}{40} = 0.375$$

Podejście bayesowskie

Z jednakowym prawdopodobieństwem możemy wylosować każdą z 3 urn – to nasza informacja o układzie.

$$g(\vec{\theta}) \longrightarrow P\{1\} = P\{2\} = P\{3\} = \frac{1}{3}$$

$$M = \int_{\Theta} L(D|\vec{\theta}) g(\vec{\theta}) d^m\theta \longrightarrow M = \sum_{i=1}^{3} P\{z|i\}P\{i\} = \frac{\frac{10}{30} + \frac{15}{30} + \frac{15}{30}}{3} = 0.44444$$

$$P\{1|z\} = P\{1\}\frac{P\{z|1\}}{M} = 0.25$$
$$P\{2|z\} = P\{2\}\frac{P\{z|2\}}{M} = 0.375$$
$$P\{3|z\} = P\{3\}\frac{P\{z|3\}}{M} = 0.375$$

- wynik identyczny jak w podejściu tradycyjnym, ale to skutek posiadania pełnej informacji o rozkładzie prawdopodobieństwa urn (P{i})

Przykład – rzut sfałszowaną monetą (wpływ początkowego fgp na wynik)

- rzut monetą daje jeden z dwóch możliwych wyników

$$\{\text{orzeł}, \text{reszka}\} = \{O, R\}$$

- jeśli moneta jest symetryczna to wyrzucenie obu zmiennych powinno być identyczne

$$P\{O\} = P\{R\} = \frac{1}{2}$$

- jeśli moneta została sfałszowana to jest niesymetryczna i prawdopodobieństwa wyrzucenia zmiennych będą różne

$$P\{O\} = p, \qquad P\{R\} = 1 - p, \qquad p \in [0, 1]$$

Nie wiemy jakie jest p, chcemy je oszacować więc użyjemy analizy bayesowskiej

$$f(p|D) = g(p)\frac{L(D|p)}{M}$$

- f(p|D) to fgp prawdopodobieństwa sukcesu (wyrzucenia orła)

- Zakładamy rozkład g(p) na podstawie naszej wiedzy.
  Ale nic nie wiemy - więc zakładamy że każdy wynik jest możliwy/dozwolony: g(p)=const

$$g(p) = \begin{cases} 1, & p \in [0, 1] \\ 0, & p \notin [0, 1] \end{cases}$$

- teraz model, który pozwala generować dane n podstawie znajomości p – to rozkład dwumianowy

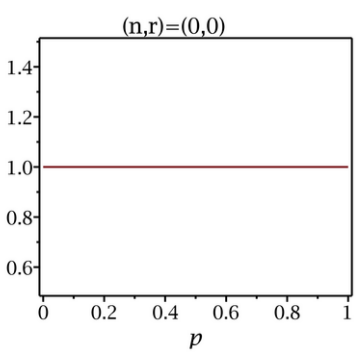
$$L(D|p) = \binom{n}{r} p^r (1-p)^{n-r}$$

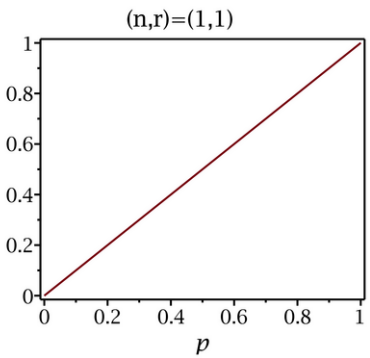
- konstruujemy fgp opisującą rozkład prawdopoodbieństwa wyrzucenia orła

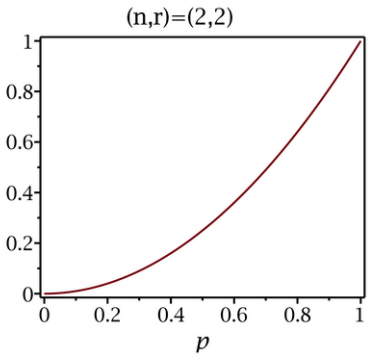
$$f(p|D) = \binom{n}{r} \frac{1}{M} p^r (1-p)^{n-r}$$

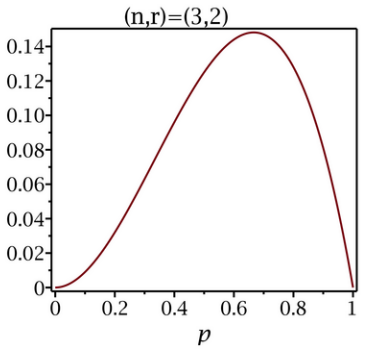
- Kolejny krok - generujemy fgp na podstawie aktualnych danych.
  Z danych pobieramy informacje (n,r) i wstawiamy do modelu, rysujemy fgp.

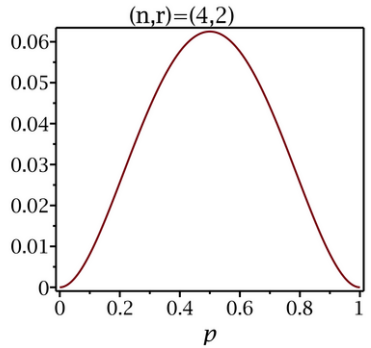
UWAGA: na rysunkach pominięto stałą normalizacji – tu nie jest istotna



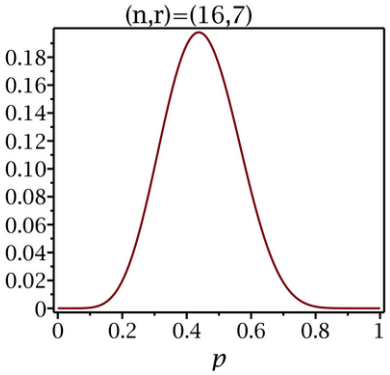


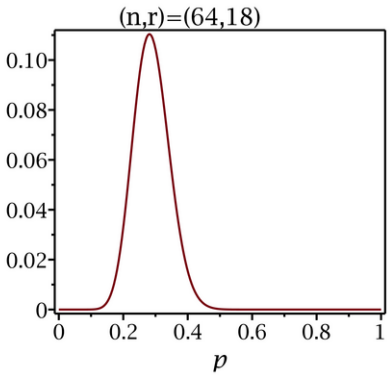


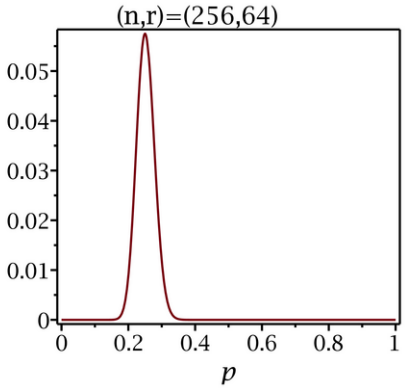


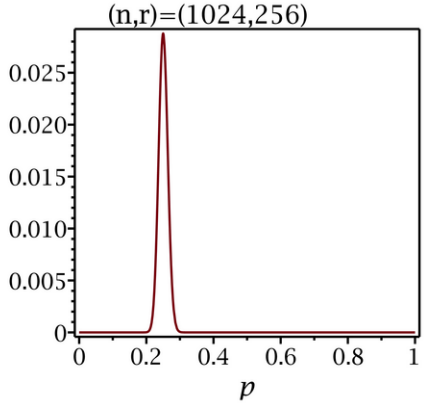


- dla małej liczby danych wyniki (fgp) znacząco ulegają zmianie w zależności od kolejnego losowania
- przedział prawdopodobnych wartości p obejmuje cały zakres









- im większa liczba danych, tym dokładniejsze oszacowanie przedziału p
- przedział niepewności p ulega zawężeniu

Zmieńmy jeszcze fgp początkowe (równomierny rozkład może nie być najlepszy).

Do modelowania wykorzystajmy znany nam rozkład beta B(α,β)

$$g(p) = B(p,\alpha,\beta) = p^{\alpha-1}(1-p)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

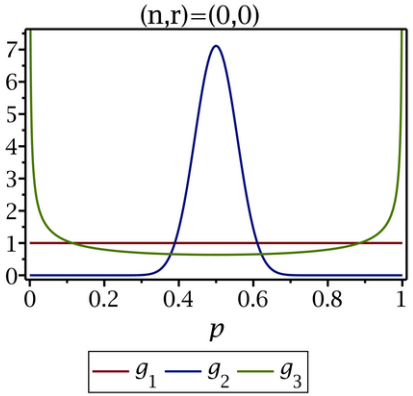
Uwaga: tutaj fgp są unormowane

3 funkcje do porównania

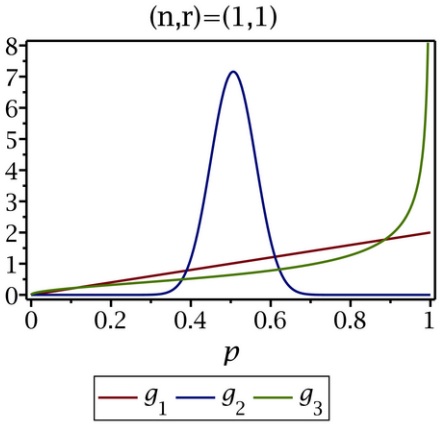
$$g_1(p) = B(p,1,1) = 1$$

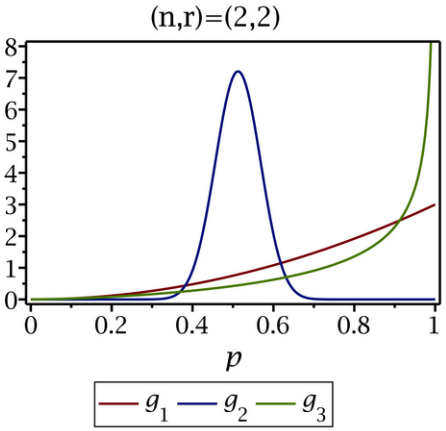
$$g_2(p) = B(p,40,40)$$

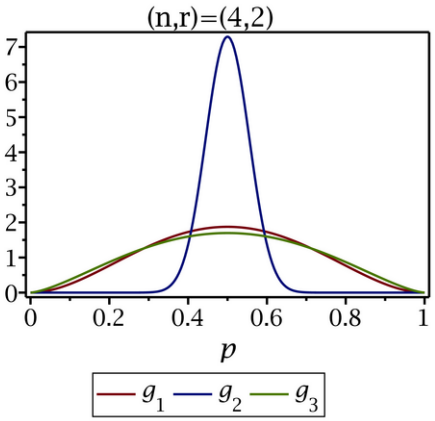
$$g_3(p) = B\left(p,\frac{1}{2},\frac{1}{2}\right)$$



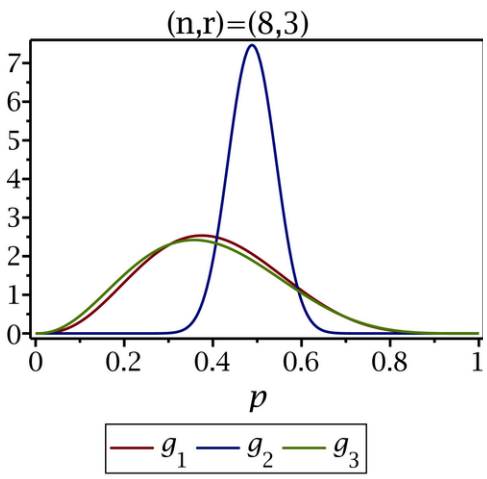
- rozkład g1: zakładamy całkowitą niewiedzę dotyczącą ewentualnego wyniku (bezpieczne założenie – niczego nie wykluczamy)
- rozkład g2: na podstawie naszego doświadczenia zakładamy, że prawdopodobieństwo wyrzucenia orła lub reszki powinno być zbliżone, silne odchylenie od tego założenia (ogon rozkładu) jest mało prawdopodobne ale go nie wykluczamy (nadal fgp>0)
- rozkład g3: podejrzewamy oszustwo, częściej będzie wypadać albo reszka albo orzeł, nie wiemy które więc asekurujemy się – dajemy duże fgp w pobliżu p=0 i p=1

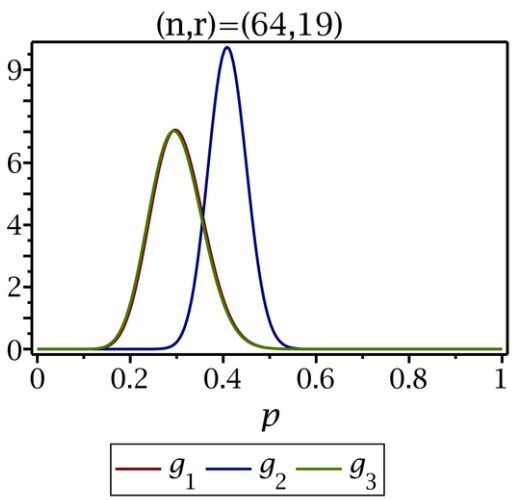


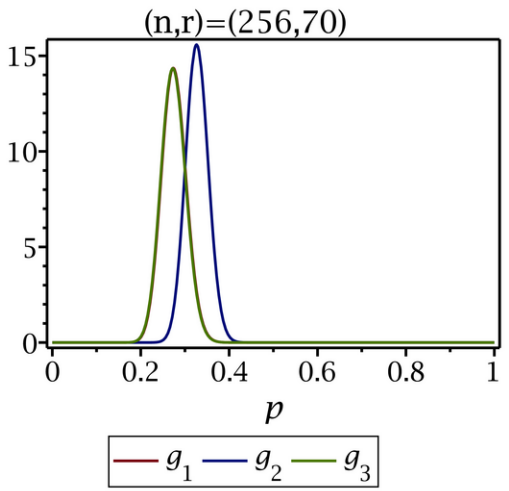


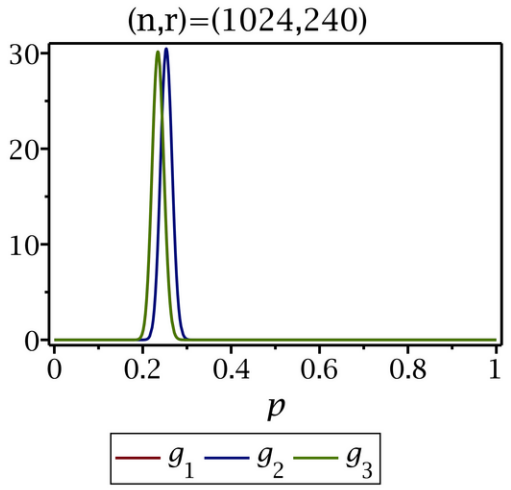


reszta wyników dla większej liczby danych









- znowu, dla małej liczby danych rozkład znacznie zmienia się i odbiega od oczekiwanego (duża rola fluktuacji)
- dla dużej liczby danych wszystkie rozkłady ewoluują taka by uzyskać ten sam końcowy rozkład
- nasza początkowa niewiedza dotycząca rozkładu p nie jest krytyczna, docelowy rozkład uzyskamy, pod warunkiem, że początkowe fgp>0 w całym obszarze
- najszybciej przesuwają się rozkłady bezpieczne (asekuranckie) – g1 i g3, gdy dopuszczaliśmy każdą ewentualność oraz możliwość fałszerstwa, najwolniej przesuwa się g2 bo w docelowym obszarze ten rozkład ma małą wartość (uznaliśmy że lokalizacja w tym obszarze jest mało prawdopodobna)

Analiza bayesowska znajduje zastosowanie w modelowaniu procesów fizycznych.
Jeśli na podstawie pomiarów (D) uda nam się znaleźć fgp f(θ|D) wówczas możemy modelować wynik procesu w funkcji farametrów θ np. jako wartość oczekiwaną pewnej funkcji z(θ)

$$\langle z \rangle = \int_\Theta z(\vec{\theta}) f(\vec{\theta}|D) d^n\theta = \int_\Theta z(\vec{\theta}) \frac{g(\vec{\theta}) L(D|\vec{\theta})}{M} d^n\theta$$

- f(θ|D) jest tylko przybliżeniem prawdziwej fgp (dokładnej nie znamy, bo np. model jest uproszczony)
- g(θ) to nasza wiedza (lub niewiedza) dotycząca rozkładu parametrów, zazwyczaj definiowana w postaci przybliżonej
- L(D|θ) funkcja wiarygodności generowana jest na podstawie danych pomiarowych, dane są obarczone błędem pomiarowym

- <z> to całościowa informacja jaką potrafimy zgromadzić w postaci przetworzonej (pojedynczej liczby) ułatwiającej analizę / podjęcie decyzji

Przykład – modelowanie struktury wewnętrznej Księżyca
(ale tego nie będziemy analizować, kto chce może sięgnąć do literatury).

Misje Apollo zostawiły zestaw sejsmografów na powierzchni Księżyca, które analizowały drgania podłoża (trzęsienia Księżyca) będących skutkiem naturalnego przemieszczania się płyt tektonicznych, uderzeń meteorytów oraz celowego rozbicia sztucznego satelity.
Analiza fal sejsmicznych, kierunków i prędkości pozwala wykalibrować wieloparametryczny model, na podstawie którego można starać się wyjaśnić (pewności nie mamy) wewnętrzną budowę Księżyca.

- K. Mosegaard, J. G. Williams, P. Lognonne, Journal of Geophysical Reaserch Atmosphere 109 (9), 2004, „Does the Moon possess a molten core? Probing the deep lunar interior using results from LLR and Lunar Prospector”
- A. Khan i inni, Geophysical Journal International 168, 243 (2007), „Joint inversion of seismic and gravity data for lunar composition and thermal state”
- R.F. Garcia i inni, Space Science Review 215, 50 (2019), „Lunar seismology: an update on interior structure models”

Estymacja wielkości charakterystycznych rozkładu: wartość oczekiwana, odchylenie standardowe
(najpierw rozważymy 1 wymiar, później uogólnimy wyniki na N>1 wymiarów)

Mimo iż naszym celem jest konstrukcja **f(θ|D)** to, pamiętajmy że określa ona pewien rozkład przestrzenny, a do analizy często lepiej używać kilku wielkości liczbowych charakterystycznych dla rozkładu.

Co jest najbardziej intersujące? położenie maksimum i szerokość rozkładu.
W maksimum f spełnia warunki (1 wymiar)

$$\left.\frac{df}{d\theta}\right|_{\theta=\theta_0} = 0 \qquad \left.\frac{d^2 f}{d\theta^2}\right|_{\theta=\theta_0} < 0$$

Ze względów praktycznych (łatwiejsze rachunki) zamiast f, posługujemy się jej logarytmem (położenie maksimum nie ulega zmianie, relacje pochodnych również)

$$W = \ln[f(\theta|D)]$$

Rozwińmy tę funkcję w szereg Taylora dla **θ=θ₀**

$$W = \underbrace{W(\theta_0)}_{=const} + \underbrace{\left.\frac{dW}{d\theta}\right|_{\theta=\theta_0}}_{=0}(\theta-\theta_0) + \frac{1}{2}\left.\frac{d^2W}{d\theta^2}\right|_{\theta=\theta_0}(\theta-\theta_0)^2 + \underbrace{O(\Delta\theta^3)}_{\approx 0}$$

$$W = W_0 + \frac{1}{2}\left.\frac{d^2W}{d\theta^2}\right|_{\theta=\theta_0}(\theta-\theta_0)^2 \approx \ln[f(\theta|D)]$$

$$f(\theta|D) = e^{W_0}\exp\left[\frac{1}{2}\left.\frac{d^2W}{d\theta^2}\right|_{\theta=\theta_0}(\theta-\theta_0)^2\right]$$

użyjmy oznaczeń

$$\left.\frac{d^2W}{d\theta^2}\right|_{\theta=\theta_0} < 0 \quad \longrightarrow \quad \left.\frac{d^2W}{d\theta^2}\right|_{\theta=\theta_0} = -\frac{1}{\sigma^2} \qquad \theta_0 = \mu$$

Uzyskaliśmy wynik, który dobrze znamy (fgp normalizujemy)

$$f(\theta|D) \;\rightarrow\; f(\theta|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right]$$

znamy też interpretację

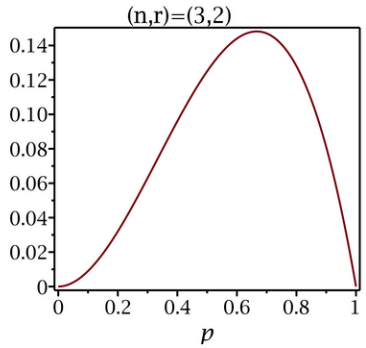
- μ to położenie maksimum rozkładu
- σ oznacza szerokość/rozmycie rozkładu, poprawny wynik mieści się w obszarze (-σ,σ) z prawdopodobieństwem ~67%

Do określenia jakości dopasowania modelu do danych wystarczy podać dwa parametry
zachowanie rozkładu w pobliżu maksimum

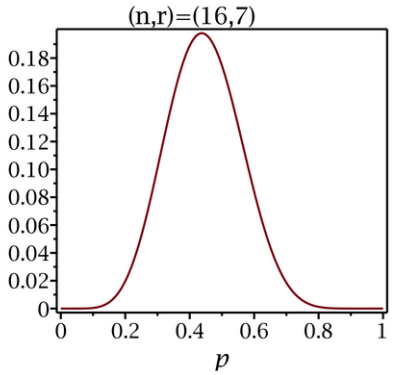
$$\theta = \mu \pm \sigma$$

Oczywiście wynik zgodnie z CTG będzie prawdziwy jeśli uwzględnimy w obliczeniach dużą liczbę danych,
to widzieliśmy też w problemie z rzutem monetą

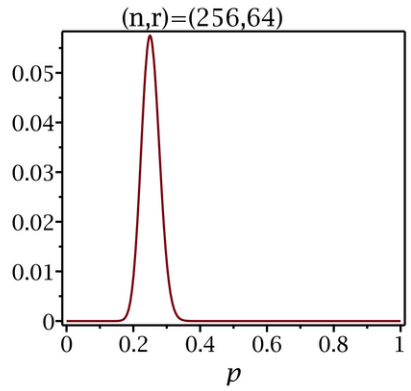
to nie przypomina rozkładu normalnego



podobny do rozkładu normalnego, ale lekko niesymetryczny (max bliższe lewej granicy)



tu rozkład gaussowski może być dobrym przybliżeniem



Przykład – szacowanie niepewności dla rzutu monetą

Weźmy model z rozkładem jednorodnym (najbardziej asekuracyjnym, najmniej wiemy o możliwym wyniku)

$$f(p|D) = \binom{n}{r} \cdot p^r (1-p)^{n-r}$$

$$W = \ln[f(p|D)] = const + r \ln p + (n-r) \ln(1-p)$$

- liczymy 1 pochodną, przyrównujemy do 0 i znajdujemy minimum

$$\frac{dW}{dp} = \frac{r}{p} - \frac{n-r}{1-p} = 0 \qquad \longrightarrow \qquad p_0 = \frac{r}{n}$$

- wynik jak w tradycyjnym podejściu określania częstości wystąpień danego zdazrenia - ale to wynik pierwotnego założenia o jednorodnym rozkładzie p (dla innego, maksimum jest przesunięte)

- liczymy wartość drugiej pochodnej w maksimum

$$\left.\frac{d^2W}{dp^2}\right|_{p=p_0} = -\frac{n}{p_0(1-p_0)} = -\frac{1}{\sigma^2}$$

- odchylenie standardowe

$$\sigma = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{var\{p(n=1)\}}{n}} = \frac{\sigma_{p(n=1)}}{\sqrt{n}}$$

- wartość licznika pod pierwiastkiem to wariancja dla pojedynczego losowania (n=1) z rozkładu dwumianowego
  → wykład 1 z rachunku prawd.

Konstrukcja funkcji wiarygodności L(D|θ) dla N>1 wymiarów

Ze względu na różnorodność problemów nie istnieje jedna metoda konstrukcji L, zależy to od przypadku

- jeśli zmienna losowa ma rozkład dyskretny, wówczas można modelować L za pomocą rozkładu wielomodalnego (podobnie do przykładu z monetą)
- model i jego parametryzacja może bezpośrednio wynikać z fizyki problemu np. liniowa zależność temperaturowa oporu metali
- metody wykorzystujące histogram – to jeden z częstych przypadków w analizie spektralnej atomowych linii widmowych
- metody bazujące na minimalizacji odległości między danymi doświadczalnymi a danymi modelowymi, w takim przypadku łatwo skonstruować fgp mające maksimum dla najmniejszej odległości

$$L(D|\vec{x}) = Ce^{-S(\vec{x})}$$

$$S(\vec{x}) = \sum_{i=1}^{N} \frac{|d_i - x_i|^{\alpha}}{\beta \cdot \sigma_i^{\alpha}}$$

- S(x) osiąga wartość minimalną w punkcie, w którym L posiada maksimum
- parametr α definiuje w jaki sposób liczona jest odległość (α=2 to odpowiednik normy eklidesowej, jeśli σ_i=const)
- parametr σ_i to niepewność pomiarowa (odchylenie standardowe),
  mała wartość σ_i wzmacnia wkład od „lepszych wyników” czyli o większej precyzji,
  natomiast duża σ_i obniża wkład „gorszych wyników” wyznaczonych z mniejszą precyzją

Przeanalizujmy ostanią metodę bazującą na minimalizacji odległości dane pomiarowe-dane modelowe

- dysponujemy danymi pomiarowymi: położenia + wartości + niepewności pomiarowe

$$D = \{y_0(x_0) \pm \sigma_0,\, y_1(x_1) \pm \sigma_1,\, \ldots, y_{N-1}(x_{N-1}) \pm \sigma_{N-1}\}$$

- tworzymy sparametryzowany model, który pozwala wygenerować dane modelowe, nieznane parametry modelu zawiera wektor α

$$y = y(x, \vec{\alpha}), \qquad \vec{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_M)$$

- w kolejnym kroku należy zaproponować rozkład zmiennej losowej y, to jest on parametryzowany załóżmy że ma ona rozkład normalny, a zmienne losowe y są niezależne

$$L(D|\vec{\alpha}) = \prod_{i=0}^{N-1} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{(y_i - y(x_i, \vec{\alpha}))^2}{2\sigma_i^2}\right]$$

uwaga 1: ponieważ wektor danych pomiarowych D jest ustalony, a elementów wektora α nie znamy, możemy potraktować fgp jako funkcję wektora α

$$L(\vec{\alpha}) = L(D|\vec{\alpha})$$

uwaga 2: iloczyn dużej liczby funkcji Gaussowskich f«1 bardzo szybko maleje do zera, dlatego znowu posłużymy się logarytmem funkcji wiarygodności

uwaga 3: chcemy znaleźć maksimum, ale numeryczne procedury optymalizacyjne zazwyczaj są konstruowane do poszukiwania minimum wartości funkcji, dlatego logarytm przemnażamy przez (-1): zamiana maksimum → minimum

$$W(\vec{\alpha}) = -\ln L(\vec{\alpha})$$

$$W(\vec{\alpha}) = -\ln L(\vec{\alpha}) \qquad \longrightarrow \qquad L(\vec{\alpha}) = e^{-W(\vec{\alpha})}$$

Interesuje nas znalezienie minimum W – to punkt o szczególnych własnościach, rozwijamy funkcję w szereg Taylora wokół tego punktu

$$\Delta\vec{\alpha} = \vec{\alpha} - \vec{\alpha}_0$$

$$W(\vec{\alpha}_0 + \Delta\vec{\alpha}) \approx W(\vec{\alpha}_0) + \sum_{i=1}^{m} \frac{\partial W}{\partial \alpha_i}\bigg|_{\vec{\alpha}=\vec{\alpha}_0} \Delta\alpha_i + \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{1}{2} \frac{\partial^2 W}{\partial\alpha_i \partial\alpha_j}\bigg|_{\vec{\alpha}=\vec{\alpha}_0} \Delta\alpha_i \Delta\alpha_j + O(|\Delta\vec{\alpha}|^3)$$

$$\sum_{i=1}^{m} \frac{\partial W}{\partial \alpha_i}\Delta\alpha_i = \vec{\nabla} W \cdot \Delta\vec{\alpha} = (\vec{\nabla} W)^T \Delta\vec{\alpha}$$

zastosowana konwencja
wektor = wektor kolumnowy

$$\vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial^2 W}{\partial\alpha_i \partial\alpha_j}\bigg|_{\vec{\alpha}=\vec{\alpha}_0} \Delta\alpha_i \Delta\alpha_j = \sum_{i=1}^{m}\sum_{j=1}^{m} \Delta\alpha_i h_{ij}(\vec{\alpha}_0) \Delta\alpha_j = \Delta\vec{\alpha}^T \boldsymbol{H}(\vec{\alpha}_0)\Delta\vec{\alpha}$$

H to macierz drugich pochodnych, macierz Hessego, hesjan

$$\boldsymbol{H} = \vec{\nabla} \otimes \vec{\nabla} = \vec{\nabla}\vec{\nabla}^T = \begin{bmatrix} \frac{\partial}{\partial\alpha_1} \\ \frac{\partial}{\partial\alpha_2} \\ \vdots \\ \frac{\partial}{\partial\alpha_m} \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial\alpha_1} & \frac{\partial}{\partial\alpha_2} & \cdots & \frac{\partial}{\partial\alpha_m} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2}{\partial\alpha_1^2} & \frac{\partial^2}{\partial\alpha_1\partial\alpha_2} & \cdots & \frac{\partial^2}{\partial\alpha_1\partial\alpha_m} \\ \frac{\partial^2}{\partial\alpha_2\partial\alpha_1} & \frac{\partial^2}{\partial\alpha_2^2} & \cdots & \frac{\partial^2}{\partial\alpha_2\partial\alpha_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\alpha_m\partial\alpha_1} & \frac{\partial^2}{\partial\alpha_m\partial\alpha_2} & \cdots & \frac{\partial^2}{\partial\alpha_m^2} \end{bmatrix}$$

$$W(\vec{\alpha}_0 + \Delta\vec{\alpha}) \approx W(\vec{\alpha}_0) + \vec{\nabla} W \cdot \Delta\vec{\alpha} + \frac{1}{2}\Delta\vec{\alpha}\boldsymbol{H}(\vec{\alpha}_0)\Delta\vec{\alpha} + O(|\Delta\vec{\alpha}|^3)$$

dostaliśmy wynik

$$W(\vec{\alpha}_0+\Delta\vec{\alpha}) \approx \underbrace{W(\vec{\alpha}_0)}_{=const} + \underbrace{\vec{\nabla}W\cdot\Delta\vec{\alpha}}_{=0} + \frac{1}{2}\Delta\vec{\alpha}\boldsymbol{H}(\vec{\alpha}_0)\Delta\vec{\alpha} + \underbrace{O(|\Delta\vec{\alpha}|^3)}_{\approx 0}$$

$$W(\vec{\alpha}) = W(\vec{\alpha}_0+\Delta\vec{\alpha}) \approx const + \frac{1}{2}\Delta\vec{\alpha}\boldsymbol{H}(\vec{\alpha}_0)\Delta\vec{\alpha}$$

wstawmy go do relacji pomiędzy W i L

$$L(\vec{\alpha}) = e^{-W(\vec{\alpha})}$$

$$L(\vec{\alpha}) \approx e^{-const}\exp\left(-\frac{\Delta\vec{\alpha}\boldsymbol{H}(\vec{\alpha}_0)\Delta\vec{\alpha}}{2}\right) = const\cdot\exp\left(-\frac{\Delta\vec{\alpha}\boldsymbol{C}^{-1}\Delta\vec{\alpha}}{2}\right)$$

$$\boldsymbol{H} = \boldsymbol{C}^{-1} \quad \longrightarrow \quad \boldsymbol{C} = \boldsymbol{H}^{-1}$$

C jest macierzą kowariancji, zawiera wariancje zmiennych i ich kowariancje (wykład - generatory)

$$\boldsymbol{C} = \begin{bmatrix} \sigma^2_{\alpha_1} & \sigma_{\alpha_1\alpha_2} & \dots & \sigma_{\alpha_1\alpha_m} \\ \sigma_{\alpha_2\alpha_1} & \sigma^2_{\alpha_2} & \dots & \sigma_{\alpha_2\alpha_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\alpha_m\alpha_1} & \sigma_{\alpha_m\alpha_2} & \dots & \sigma^2_{\alpha_m} \end{bmatrix}$$

współczynnik korelacji 2 parametrów/zmiennych

$$r_{ij} = \frac{\sigma_{\alpha_i\alpha_j}}{\sqrt{\sigma^2_{\alpha_i}\sigma^2_{\alpha_j}}}$$

- jeśli znajdziemy minimum W (maksimum L) oraz wszystkie drugie pochodne (hesjan) to po odwróceniu macierzy H dostaniemy niepewności szacowanych parametrów α
- **pozostaje pytanie: jak znaleźć minimum W (maksimum L)?**

Minimalizacja funkcji W (szukamy maksimum L)

$$L(D|\vec{\alpha}) = \prod_{i=0}^{N-1} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{(y_i - y(x_i,\vec{\alpha}))^2}{2\sigma_i^2}\right]$$

$$W(\vec{\alpha}) = -\ln\, L(\vec{\alpha})$$

wstawiamy L do wzoru na W i zapisujemy jawną postać W

$$W = \underbrace{-\ln\left(\prod_{i=0}^{N-1}\sigma_i^{-1}(\sqrt{2\pi})^{-1}\right)}_{=const} + \frac{1}{2}\underbrace{\sum_{i=0}^{N-1}\frac{[y_i - y(x_i;\vec{\alpha})]^2}{\sigma_i^2}}_{\chi^2}$$

tylko drugi wyraz zależy od parametrów α – więc minimalizację należy przeprowadzić w oparciu o jego wartość

$$\chi^2(\vec{\alpha}) = \sum_{i=1}^{N-1}\frac{[y_i - y(x_i;\vec{\alpha})]^2}{\sigma_i^2}$$

- dostaliśmy prosty wzór, który formalnie przedstawia metodę aproksymacji średniokwadratowej
- jeśli mamy tylko jeden parametr to minimalizację możemy wykonać analitycznie przekształacjąc wzór, dla większej liczby parametrów używamy gotowych procedur numerycznych
- kolejna kwestia wiąże się ze sposobem określania niepewności σ – rozważymy dwa przypadki:

  1) dane przetworzone w postaci histogramu (najczęściej spotykany przypadek, interesują nas tylko dane)

  2) „surowe” nieprztworzone dane (rzadko spotykamy – stosowany w bardziej zaawansowanej analizie, kalibracja aparatury, etc.)

1 - szacowanie parametrów α dla danych przetworzonych w postaci histogramu

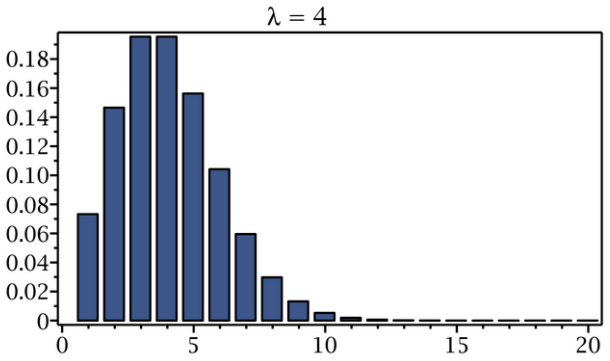
Rozważamy standardowy proces stochastyczny – rozpad promieniotwórczy.
fpg tego procesu ma rozkład dyskretny – to **rozkład Poissona**

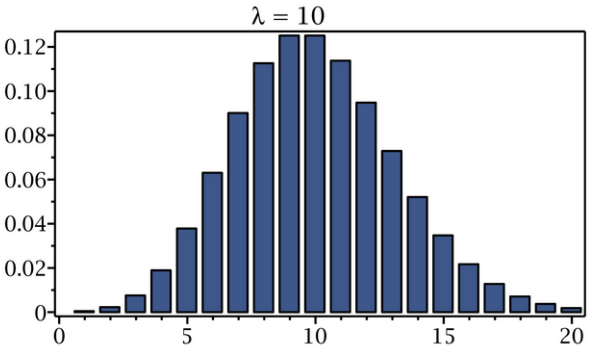
$$f(n;\lambda) = \frac{\lambda^n}{n!} e^{-\lambda}, \qquad n = 1, 2, \ldots$$

λ będzie parametrem modelu – szukamy jej wartości

Niech N oznacza liczbę zliczeń rejestrowaną przez licznik Geigera-Mullera w równych odstępach czasu.
Histogram tworzymy jako liczbę wystąpień kolejnych wartości n=1,2,3,4,...

rozkład Poissona





Wariancja i odchylenie standardowe rozkładu (wykład 1) – poza przypadkiem, gdy N=0:

$$\sigma_n^2 = n \qquad \longrightarrow \qquad \sigma_n = \sqrt{n}$$

Dane pomiarowe to ilości zarejestrowanych przypadków dla n=1,2,3,4,....

$$\begin{aligned} D &= \{n_1 \pm \sigma_1, n_2 \pm \sigma_2, \ldots, n_N \pm \sigma_N\} \\ &= \{n_1 \pm \sqrt{n_1}, n_2 \pm \sqrt{n_2}, \ldots, n_N \pm \sqrt{n_N}\} \end{aligned}$$

określamy całkowitą liczbę zliczeń

$$n = \sum_{i=1}^{N} n_i$$

oraz wartości modelowe $\mathbf{m_i}$

$$m_i = n \cdot f(i;\lambda)$$

konstruujemy funkcję celu

$$\chi^2(\lambda) = \sum_{i=1}^{N} \frac{[n_i - n\, f(i;\lambda)]^2}{n_i}$$

W przypadku zmiennej ciągłej, dzielimy obszar na podprzedziały i każdemu z nich przypisujemy zarejestrowaną liczbę zliczeń w danym przedziale

$$n_i \rightarrow x_i < x \leq x_i + \Delta x_i$$

prawdopodobieństwo wylosowania zmiennej z danego przedziału określamy na podstawie fgp rozkładu modelowego i jego dystrubuanty

$$p_i = F(x_i + \Delta x_i; \lambda) - F(x_i; \lambda) = \int_{x_i}^{x_i + \Delta x_i} f(x, \lambda) dx$$

modelową liczbę zliczeń w danym przedziale określimy jako iloczyn

$$m_i = n \cdot p_i$$

Uwaga: jeśli rozkład rejestrowanej zmiennej losowej rozciąga się poza obszar objęty histogramem, wówczas musimy dokonać renormalizacji danych modelowych, n dotyczy całego rozkładu (nie uciętego jak w przypadku histogramu)

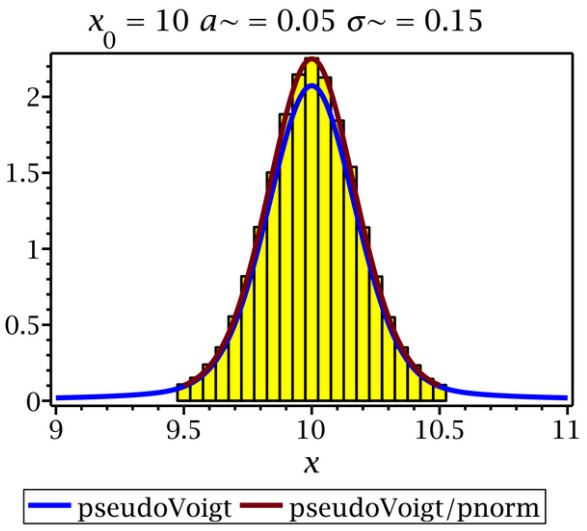
$$p_{norm} = F(x_{max}) - F(x_{min}) \leq 1$$

i dokonujemy renormalizacji wyniku modelowego

$$m_{i,norm} = n \cdot \frac{p_i}{p_{norm}}$$

$$\chi^2(\lambda) = \sum_{i=1}^{N} \frac{\left[n_i - n \frac{p_i}{p_{norm}} \right]^2}{n_i}$$

3-parametryczny rozkład (pseudo)Voigta (histogram – rozkład obcięty, x – energia)



2 – szacowanie parametrów dla danych nie przetworzonych

Jako przykład może posłużyć rejestrowanie pojedycznych impulsów w spektrometrze, każdemu impulsowi przypisujemy zmierzoną energię

$$D = \{E_1, E_2, \ldots, E_N\}$$

- liczba surowych danych może być duża: 10^3-10^8
- natomiast histogram zawiera przeciętnie: 10^1-10^3 przedziałów

Ponieważ kolejne zarejestrowane zmienne losowe są niezależne, możemy utworzyć fgp rozkładu N-wymiarowego w postaci iloczynu rozkładu pojedynczej zmiennej

$$x_i \sim Dist\{f(x, \vec{\alpha})\} \qquad \longrightarrow \qquad D = \{x_1, x_2, \ldots, x_N\}$$

$$L(D, \vec{\alpha}) = \prod_{i=1}^{N} f(x_i, \vec{\alpha})$$

- rozkład zmiennej dowolny, wynika z fizyki problemu

wprowadzamy funkcję pomocniczą W

$$W = -\ln L(D, \vec{\alpha}) = -\sum_{i=1}^{N} \ln f(x_i, \vec{\alpha})$$

- minimalizację możemy wykonać działając jedynie na wartościach W – metoda SIMPLEX (Neldera-Meada)
- metody gradientowe wymagają liczenia gradientu (który w minimum znika)

$$\frac{\partial W}{\partial \alpha_i} = -\sum_{i=1}^{N} \frac{\ln f(x_i, \vec{\alpha})}{\partial \alpha_i} = -\sum_{i=1}^{N} \frac{1}{f(x_i, \vec{\alpha}_i)} \frac{\partial f(x_i, \vec{\alpha})}{\partial \alpha_i}$$

- ponieważ szacujemy wartości parametrów α, przydadzą nam się też ich odchylenia standardowe, liczymy **hesjan** (jeśli używamy metody Newtona w minimalizacji to hesjan i tak jest potrzebny)

$$\boldsymbol{H} = \vec{\nabla}_{\vec{\alpha}} \otimes \vec{\nabla}_{\vec{\alpha}} W$$

$$h_{ij} = \frac{\partial^2 W}{\partial\alpha_i \partial\alpha_j} = \frac{\partial}{\partial\alpha_j}\left(-\sum_{i=1}^{N} \frac{1}{f(x_i,\vec{\alpha})} \frac{\partial f(x_i,\vec{\alpha})}{\partial\alpha_i}\right)$$

$$= -\sum_{i=1}^{N}\left(-\frac{1}{f^2(x_i,\vec{\alpha})} \frac{\partial f(x_i,\vec{\alpha})}{\partial\alpha_i} \frac{\partial f(x_i,\vec{\alpha})}{\partial\alpha_j} + \frac{1}{f(x_i,\vec{\alpha})} \frac{\partial^2 f(x_i,\vec{\alpha})}{\partial\alpha_i \partial\alpha_j}\right)$$

Uwagi:

- minimalizację wykonujemy przy użyciu gotowych procedur numerycznych w sposób iteracyjny, w każdej iteracji liczymy wartość funkcji i zazwyczaj też gradient, metoda wykorzystująca histogram będzie działać szybciej bo sumowanie wykonujemy na mniejszej liczbie danych (10-1000 razy mniej)
- w obu przypadkach (histogram i surowe dane) musimy policzyć hesjan – bo po odwróceniu daje macierz kowariancji, z niej wyciągamy odchylenia standardowe i współczynniki korelacji
- w przypadku rozkładu, dla którego fgp opisywana jest jawną zależnością funkcyjną (funkcje elementarne/specjalne) z różniczkowaniem po parametrze α nie ma problemu
- jeśli fgp jest rozkładem złożonym (np. Voigt = splot Gaussa i Cauchy'ego) i nie znamy zależności funkcyjnej, można go uprościć (przybliżyć/aproksymować) inną funkcją, której wartości łatwo obliczyć numerycznie (np. pseudoVoigt)
- różniczkowanie wykonujemy zazwyczaj, przy użyciu programów do obliczeń symbolicznych typu Maple/Mathematica

Przykład – analiza spektralna linii emisyjnych atomów

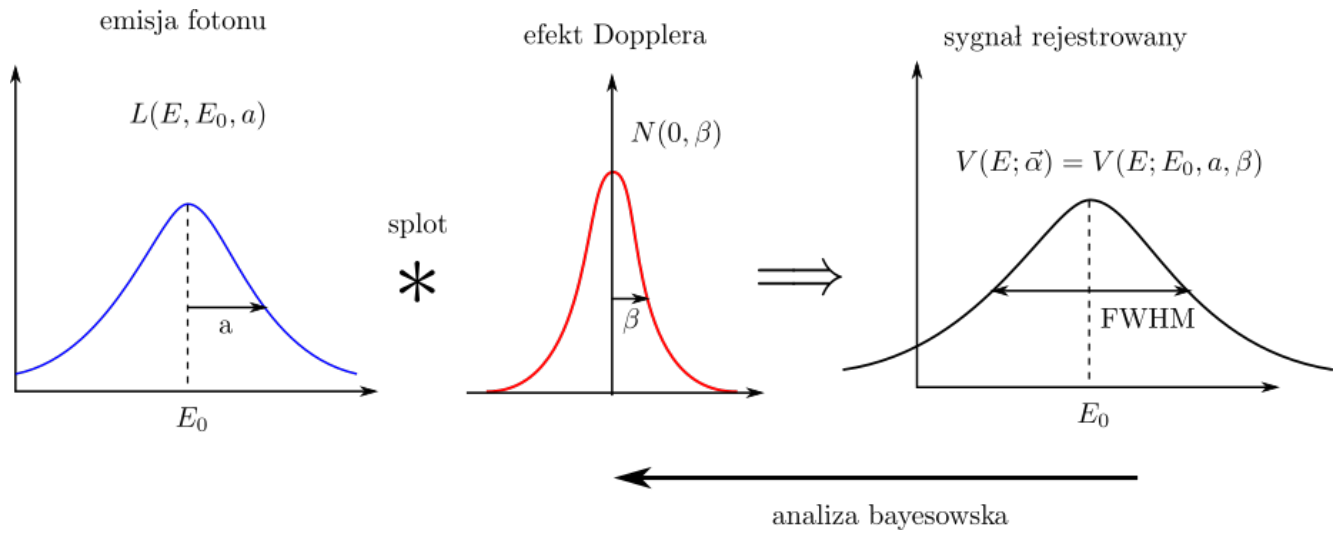
W astrofizyce jednym z niewielu dostępnych narzędzi eksperymentalnych są badania spektroskopowe.
Atomy na powierzchnii (odległej) gwiazdy emitują fotony w szerokim zakresie energii,
analiza widm spektralnych dostarcza więc informacji nt:

- składu pierwiastkowego (natężenie linii)
- odległości (przesunięcie ku czerwieni)
- temperatury powierzchnii (ruch termiczny atomów – efekt Dopplera)

Jak wykorzystać analizę bayesowską w modelowaniu widm?

- fgp opisująca pojedynczy pik w widmie ma rozkład Voigta (splot rozkładu normalnego i Cauchy'ego)
  ten rozkład to nasz model

$$f_V(E,a,\beta) = \int_{-\infty}^{\infty} \left[\frac{1}{\pi}\frac{a}{a^2+(E-E')^2}\right]\left[\frac{1}{\beta\sqrt{2\pi}}\exp\left(-\frac{E'^2}{2\beta^2}\right)\right] dE'$$



FWHM=
full width at half maximum

nasz cel to wydobycie z widma: $a \pm \sigma_a, \quad E_0 \pm \sigma_{E_0}, \quad \beta \pm \sigma_\beta$

- stosowanie rozkładu Voigta w obliczeniach (także analitycznych) stanowi problem, jego fgp jest splotem, ma postać analityczną, ale nieco skomplikowaną

$$f_V(E,\Gamma_G,\Gamma_L) = \frac{1}{\sqrt{\pi}\left(\frac{\Gamma_G}{2\sqrt{\ln 2}}\right)} K\left(\frac{E}{\frac{\Gamma_G}{2\sqrt{\ln 2}}}, \frac{\Gamma_L}{\Gamma_G}\sqrt{\ln 2}\right), \qquad \Gamma_L = 2a, \qquad \Gamma_G = 2\sqrt{2\ln 2}\beta$$

$$K(x,y) = Re\{w(x+iy)\} \qquad w(z) = \exp(-z^2)erfc(-iz)$$

- erfc() to dopełnienie funkcji błędu

to dopiero fgp, a trzeba jeszcze wyznaczyć dystrybuantę....

- stosujemy wzór przybliżony tzw. rozkład pseudoVoigta
  źródło parametryzacji: T.Ida et al., J. Appl. Crystal. 33, 1311-1316 (2000)

$$f_{V_p}(E,E_0,\gamma_G,\gamma_L) = (1-\eta)N(E-E_0,\gamma_G) + \eta L(E-E_0,\gamma_L)$$

$$\eta \in [0,1] \qquad \gamma_G = \frac{\Gamma}{2\sqrt{\ln 2}} \qquad \gamma_L = \frac{\Gamma}{2} \qquad \Gamma_L = 2a, \qquad \Gamma_G = 2\sqrt{2\ln 2}\beta$$

$$\eta = 1.36603\left(\frac{\Gamma_L}{\Gamma}\right) - 0.47719\left(\frac{\Gamma_L}{\Gamma}\right)^2 + 0.11116\left(\frac{\Gamma_L}{\Gamma}\right)^3$$

$$\Gamma = \left(\Gamma_G^5 + 2.69269\Gamma_G^4\Gamma_L + 2.42843\Gamma_G^3\Gamma_L^2 + 4.47163\Gamma_G^2\Gamma_L^3 + 0.07842\Gamma_G\Gamma_L^4 + \Gamma_L^5\right)^{\frac{1}{5}}$$

Rozkład pseudoVoigta łatwo scałkujemy (dystrybuanta). Problem może stanowić policzenie pochodnych f_{Vp} względem parametrów (parametry a i β są nie tylko w rozkładach składowych, ale także w parametrze pomocniczym η)
- różniczkowanie lepiej zlecić programowi Maple/Mathematica

- do celów testowych możemy użyć: (1) danych eksperymentalnych lub (2) wygenerować je
  wybieramy opcję nr (2)

$$
\begin{aligned}
&ustaw: \quad E_0, a, \beta, \eta \\
&for(i = 0; i < N; i++)do \\
&\quad U_1 \sim U(0,1) \\
&\quad if(U_1 < \eta)then \\
&\qquad E \sim L(E - E_0, a) \\
&\quad else \\
&\qquad E \sim N(E - E_0, \beta) \\
&\quad end\,if \\
&end\,do
\end{aligned}
$$

- po wygenerowaniu wyników decydujemy czy analizę bayesowską stosujemy dla surowych danych czy dla histogramu, wykonujemy minimalizację funkcji wiarygodności i liczymy hesjan/macierz kowariancji

wyniki dla parametrów: $E_0 = 10, \quad a = 0.25, \quad \beta = 1.25, \quad N = 10^6$

histogram: M=14 przedziałów

$E_0 = 9.99725 \pm 0.00109 \qquad a = 0.248571 \pm 0.00202 \qquad \beta = 1.25560 \pm 0.00190$

współczynniki korelacji

$r_{E_0,a} = -0.0006 \qquad r_{E_0,\beta} = 0.0024 \qquad r_{a,\beta} = -0.8655$

dane nieprzetworzone:

$E_0 = 9.99757 \pm 0.00108 \qquad a = 0.247101 \pm 0.00084 \qquad \beta = 1.25682 \pm 0.00116$

$r_{E_0,a} = -0.0012 \qquad r_{E_0,\beta} = 0.0016 \qquad r_{a,\beta} = -0.5966$

wnioski:

- oszacowanie wartości parametrów rozkładu jest dobre, są bliskie wartościom dla jakich zostały wygenerowane dane (ale pamiętajmy że nie było szumu tła, który pogorszyłby jakość oszacowania)
- niepewności są bardzo małe w obu przypadkach, ale wynika to z faktu że użyliśmy dużej liczby danych co daje małe fluktuacje statystyczne
- brak korelacji pomiędzy położeniem piku (E_0) a dwoma pozostałymi parametrami (a i β)
- parametry a i β są silnie ujemnie skorelowane: jeśli jeden maleje to drugi rośnie i vice versa, to wynika z faktu że rozkład jest złożeniem rozkładów normalnego i Cauchy'ego

Łączenie danych – metoda podstawowa

- w przypadku analizy danych po zakończeniu eksperymentu, czas poświęcamy na wykonanie analizy ma znaczenie drugorzędne
- inaczej jest w przypadku, gdy eksperyment jest w toku i konieczne jest wykonywanie analizy na bieżąco, po to aby móc reagować np. na zmianę warunków otoczenia, które zaburzają pomiar
- gromadzenie danych w sposób ciągły powoduje, iż prowadzenie analizy bayesowskiej w standardowy sposób staje się bezcelowe, wraz ze wzrostem liczby danych wydłużałby się czas potrzebny na jej wykonanie – a reagować chcemy natychmiast
- minimalizacja czasu potrzebnego na wykonanie analizy dla powiększonego zbioru danych możliwa jest dzięki użyciu prostej **metody łączenia danych**

Załóżmy, że mamy dwie serie danych eksperymentalnych, dla których używamy tego samego modelu dla funkcji wiarygodności (celem jest wyznaczenie parametrów α)

$$\begin{aligned} D_1 \quad &\rightarrow \quad L(D_1;\vec{\alpha}) = L_1(\vec{\alpha}) \quad \rightarrow \quad W_1(\vec{\alpha}) = -\ln L_1(\vec{\alpha}) \\ D_2 \quad &\rightarrow \quad L(D_2;\vec{\alpha}) = L_2(\vec{\alpha}) \quad \rightarrow \quad W_2(\vec{\alpha}) = -\ln L_2(\vec{\alpha}) \end{aligned}$$

zakładając że oba zbiory danych są niezależne możemy skonstruować wypadkową funkcję L_{12}

$$L_{12}(\vec{\alpha}) = L_1(\vec{\alpha})L_2(\vec{\alpha}) \quad \rightarrow \quad W_{12}(\vec{\alpha}) = W_1(\alpha) + W_2(\alpha)$$

zastosowanie kolejnych kroków procedury minimalizacyjnej (kilka slajdów wstecz) prowadzi do relacji

$$\chi^2_{12}(\vec{\alpha}) = \chi^2_1(\vec{\alpha}) + \chi^2_2(\vec{\alpha})$$

minimalizacja obu funkcji składowych z osobna daje dwa wektory α (bo dane są różne)

$$\min \chi^2_k(\vec{\alpha}) \rightarrow \vec{\alpha}_k, \qquad k = 1, 2$$

Każdą z tych funkcji rozwijamy w szereg Taylora punkcie stanowiącym jej minimum do wyrazów rzędu 2, (gradient w minimum znika) – punkty α_1 i α_2 są leżą blisko siebie

$$\chi^2_{12}(\vec{\alpha}) \approx const + (\vec{\alpha} - \vec{\alpha}_1)^T \boldsymbol{C}_1^{-1} (\vec{\alpha} - \vec{\alpha}_1) + (\vec{\alpha} - \vec{\alpha}_2)^T \boldsymbol{C}_2^{-1} (\vec{\alpha} - \vec{\alpha}_2)$$

C_1, C_2 - macierze kowariancji dla obu zbiorów danych

Ponieważ oba zbiory możemy połączyć w jeden, więc możemy powyższe równanie zapisać w zwartej postaci

$$\chi^2_{12}(\vec{\alpha}) \approx const + (\vec{\alpha} - \vec{\alpha}_{12})^T \boldsymbol{C}_{12}^{-1} (\vec{\alpha} - \vec{\alpha}_{12})$$

policzmy hesjany obu wyrażeń i przyrównajmy je do siebie

$$\vec{\nabla} \otimes \vec{\nabla} \chi^2_{12} = \boldsymbol{C}_{12}^{-1} = \boldsymbol{C}_1^{-1} + \boldsymbol{C}_2^{-1} \qquad \longrightarrow \qquad \boldsymbol{C}_{12}^{-1} = \boldsymbol{H}_{12} = \boldsymbol{H}_1 + \boldsymbol{H}_2$$

dostajemy relację pomiędzy macierzami kowariancji

$$\boldsymbol{C}_{12}^{-1} = \boldsymbol{C}_1^{-1} + \boldsymbol{C}_2^{-1}$$

$$\boldsymbol{C}_{12} = (\boldsymbol{C}_1^{-1} + \boldsymbol{C}_2^{-1})^{-1} = (\boldsymbol{H}_1 + \boldsymbol{H}_2)^{-1}$$

Połączenie obu zbiorów danych zmienia położenie minimum (nie będzie średnią z obu znalezionych wartości). Postępujemy standardowo - liczymy gradient nowej funkcji i przyrównujemy wynik do 0.

$$\vec{\nabla} \chi^2_{12} = \vec{\nabla} \chi^2_1 + \vec{\nabla} \chi^2_2 = 2\boldsymbol{C}_1^{-1}(\vec{\alpha} - \vec{\alpha}_1) + 2\boldsymbol{C}_2^{-1}(\vec{\alpha} - \vec{\alpha}_2) = \vec{0}$$

$$(\boldsymbol{C}_1^{-1} + \boldsymbol{C}_1^{-1})\vec{\alpha} = \boldsymbol{C}_1^{-1}\vec{\alpha}_1 + \boldsymbol{C}_2^{-1}\vec{\alpha}_2$$

$$\vec{\alpha} = (\boldsymbol{C}_1^{-1} + \boldsymbol{C}_1^{-1})^{-1}[\boldsymbol{C}_1^{-1}\vec{\alpha}_1 + \boldsymbol{C}_2^{-1}\vec{\alpha}_2]$$

$$\vec{\alpha} = \boldsymbol{C}_{12}[\boldsymbol{C}_1^{-1}\vec{\alpha}_1 + \boldsymbol{C}_2^{-1}\vec{\alpha}_2]$$

przykład

$$y_1 \pm \sigma_1, \quad y_2 \pm \sigma_2$$

$$\sigma_{12}^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-1}$$

$$y_{12} = \frac{\frac{y_1}{\sigma_1^2} + \frac{y_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

Łączenie danych – metoda Kalmana

W podstawowej metodzi łączenia danych zakładaliśmy, że oba zbiory danych D_1 i D_2 są liczne.
Teraz rozważymy bardziej skrajny przypadek: D_1 zawiera n-1 danych, a D_2 tylko jeden wynik

$$D_1 = \{\vec{y}_1(x_1) \pm \vec{\sigma}_1, \vec{y}_2(x_2) \pm \vec{\sigma}_2, \ldots, \vec{y}_{n-1}(x_{n-1}) \pm \vec{\sigma}_{n-1}\}$$

$$D_2 = \{\vec{y}_n(x_n) \pm \vec{\sigma}_n\}$$

uwagi:

- zakładamy bardziej ogólną sytuację tj. jeden punkt pomiarowy może zawierać więcej niż jeden wynik → wektor y
- tu zakładamy że wektor D_1 jest wektorem startowym tj. takim, od którego rozpoczynamy całą procedurę, ale jak zobaczymy, nie musi tak być, może to być już o wiele dalszy etap zbierania danych

Funkcję celu dla zestawu D_1 znamy, zawiera informację o przybliżonych wartościach:
wektora α_{n-1} oraz macierz kowariancji $\mathbf{C}_{n-1}$ – dostajemy je w wyniku zastosowania pierwotnej metody na surowych danych lub z zastosowania filtru Kalmana (dane już połączone)

$$\chi^2_{n-1}(\vec{\alpha}) \approx (\vec{\alpha} - \vec{\alpha}_{n-1})^T \boldsymbol{C}^{-1}_{n-1} (\vec{\alpha} - \vec{\alpha}_{n-1})$$

natomiast dla zestawu D_2 stosujemy standardową procedurę

$$\chi^2_{new}(\vec{\alpha}) = \left(\vec{y}_n - \vec{f}(x_n, \vec{\alpha})\right)^T \boldsymbol{R_n}^{-1} \left(\vec{y}_n - \vec{f}(x_n, \vec{\alpha})\right)$$

- $\mathbf{R}_n$ to macierz kowariancji. wartość α nie jest tu znana

tworzymy funkcję celu dla obu zestawów danych (n danych – stąd dolny wskaźnik)

$$\chi^2_n(\vec{\alpha}) = \chi^2_{n-1}(\vec{\alpha}) + \chi^2_{new}(\vec{\alpha})$$

$$\chi^2_n(\vec{\alpha}) = (\vec{\alpha} - \vec{\alpha}_{n-1})^T \boldsymbol{C}^{-1}_{n-1} (\vec{\alpha} - \vec{\alpha}_{n-1}) + \left(\vec{y}_n - \vec{f}(x_n, \vec{\alpha})\right)^T \boldsymbol{R_n}^{-1} \left(\vec{y}_n - \vec{f}(x_n, \vec{\alpha})\right)$$

Funkcję celu wyraziliśmy w postaci sumy

$$\chi^2_n(\vec{\alpha}) = (\vec{\alpha} - \vec{\alpha}_{n-1})^T \boldsymbol{C}^{-1}_{n-1}(\vec{\alpha} - \vec{\alpha}_{n-1}) + \left(\vec{y}_n - \vec{f}(x_n, \vec{\alpha})\right)^T \boldsymbol{R_n}^{-1} \left(\vec{y}_n - \vec{f}(x_n, \vec{\alpha})\right)$$

zapisujemy ją w zwięzłej postaci z nowym (nieznanym) przybliżeniem wektora α_n

$$\chi^2_n(\vec{\alpha}) = (\vec{\alpha} - \vec{\alpha}_n)^T \boldsymbol{C}^{-1}_{n}(\vec{\alpha} - \vec{\alpha}_n)$$

W pierwszym równaniu (suma) mamy niejawną zależność od α, ale możemy rozwinąć funkcję wektorową f() w szereg Taylora w punkcie, który znamy: α_{n-1}. Dla składowej i-tej mamy

$$f_i(x_n, \vec{\alpha}) = \underbrace{f_i(x_n, \vec{\alpha}_{n-1})}_{=const} + \vec{\nabla}_{\vec{\alpha}} f_i(x_n \vec{\alpha}) \cdot (\vec{\alpha} - \vec{\alpha}_{n-1}) + \ldots$$

co możemy uogólnić dla całego wektora

$$\vec{f}(x_n, \vec{\alpha}) = \underbrace{\vec{f}(x_n, \vec{\alpha}_{n-1})}_{=\vec{const}} + \boldsymbol{A}_{n-1}(\vec{\alpha} - \vec{\alpha}_{n-1}) + \ldots$$

macierz **A** zawiera pochodne f_i względem α_j

$$[\boldsymbol{A}_{n-1}]_{ij} = \left.\frac{\partial f_i(x_n, \vec{\alpha})}{\partial \alpha_j}\right|_{\vec{\alpha}=\vec{\alpha}_{n-1}}$$

podstawiamy rozwinięcie do wyrażenia z sumą i wyznaczamy macierz kowariancji $\mathbf{C}_n$

Podstawiamy rozwinięcie do wyrażenia z sumą i wyznaczamy macierz kowariancji $\mathbf{C}_n$

$$\begin{aligned}\chi^2_n(\vec{\alpha}) &= (\vec{\alpha}-\vec{\alpha}_n)^T \boldsymbol{C}_n^{-1}(\vec{\alpha}-\vec{\alpha}_n) \\ &= (\vec{\alpha}-\vec{\alpha}_{n-1})^T \boldsymbol{C}_{n-1}^{-1}(\vec{\alpha}-\vec{\alpha}_{n-1}) \\ &+ \left(\vec{y}_n - \vec{f}(x_n,\vec{\alpha}_{n-1}) - \boldsymbol{A}_{n-1}(\vec{\alpha}-\vec{\alpha}_{n-1})\right)^T \boldsymbol{R_n}^{-1} \left(\vec{y}_n - \vec{f}(x_n,\vec{\alpha}_{n-1}) - \boldsymbol{A}_{n-1}(\vec{\alpha}-\vec{\alpha}_{n-1})\right)\end{aligned}$$

Liczymy hesjany obu stron wyrażenia

$$\vec{\nabla}\otimes\vec{\nabla}\chi^2_n = \boldsymbol{C}_n^{-1} = \boldsymbol{C}_{n-1}^{-1} + \boldsymbol{A}_{n-1}^T \boldsymbol{R}_n^{-1} \boldsymbol{A}_{n-1}$$

czyli macierz kowariancji dla połączonych danych ma postać

$$\boldsymbol{C}_n = \left(\boldsymbol{C}_{n-1}^{-1} + \boldsymbol{A}_{n-1}^T \boldsymbol{R}_n^{-1} \boldsymbol{A}_{n-1}\right)^{-1}$$

Nowe oszacowanie wektora parametrów α dostaniemy licząc gradient i przyrównując go do 0

$$\vec{\nabla}\chi^2_n = \vec{0}$$

$$\vec{\alpha}_n = \boldsymbol{C}_n \left(\boldsymbol{C}_{n-1}^{-1}\vec{\alpha}_{n-1} + \boldsymbol{A}_{n-1}^T \boldsymbol{R}_n^{-1}\left[\vec{y}_n - \vec{f}(x_n,\vec{\alpha}_{n-1}) + \boldsymbol{A}_{n-1}\vec{\alpha}_{n-1}\right]\right)$$

Uwagi:

- dostaliśmy wyrażenia na α_n i $\mathbf{C}_n$, które wykorzystujemy na dwa sposoby:
  (1) dokonujemy natychmiastowej zmiany parametrów eksperymentu i/lub
  (2) używamy ich w kolejnym cyklu łączenia danych, gdy te się pojawią

- para (α_n, $\mathbf{C}_n$) stanowi naszą wiedzę o aktualnym rozkładzie parametrów,
  co jest istotne w analizie bayesowskiej

- w każdym kolejnym cyklu łączenia danych nasza wiedza o rozkładzie parametrów
  powiększa się, więc wyniki będą coraz dokładniejsze a niepewności mniejsze

- należy pamiętać, że celem stosowania metody Kalmana jest jej duża wydajność,
  przetwarzanie danych polega na odwracaniu/mnożeniu macierzy o liczbie wierszy i kolumn
  rzędu 10x10 – wynik otrzymujemy natychmiast, natomiast w metodzie podstawowej operacje
  wykonywane są na zbiorach kilkudziesięciu/kilkuset tysięcy co prowadzi do wydłużenia
  czasu obliczeniowego

Zastosowania:

- po raz pierwszy użyto podczas lądowania amerykańskich astronautów na Księżycu w 1960 roku,
  aby komputer pokładowy na bieżąco/natychmiast reagował na zmieniające się warunki

- sterowanie trajektorą satelitów, te poruszajce się w niewielkiej odległości od Ziemi (~1000 km)
  oddziałują z rozrzedzoną atmosferą, ich ruch jest więc zaburzany (fluktuacje) i konieczne jest
  wprowadzanie poprawek do ich aktualnej trajektorii, sterowanie większością satelitów (kilka tysięcy)
  musi też się odbywać automatycznie – komputery sterujące potrzebują do tego informacji

- starty rakiet kosmicznych są w pełni zautomatyzowane, dane zbierane są z kilkuset czujników,
  po ich przetworzeniu podejmowane są decyzje – 5 komputerów głosuje, liczba nieparzysta eliminuje
  możliwość braku podjęcia decyzji

- sterowanie pociskami samonaprowadzającymi w wojsku – w trakcie lotu pocisk oddziałuje z atmosferą
  (podmuchy wiatru) więc aby trafił do celu konieczna jest automatyczna kontrola lotu, ponieważ pocisk
  porusza się z dużą prędkością (1000-5000 km/s) a lot trwa maksymalnie kilka minut, komputer sterujący
  musi na bieżąco zbierać dane, przetwarzać je (filtr Kalmana) i podejmować decyzje