

Analiza danych

Mariusz Przybycień

Wydział Fizyki i Informatyki Stosowanej
Akademia Górniczo-Hutnicza

Wykład 5

- Przykład: W celu wyznaczenia wydajności detektora, umieszczamy na wiązce (jeden za drugim) dwa detektory o nieznanach wydajnościach p_A i p_B . Niech μ oznacza nieznaną liczbę cząstek padających na detektory, a n_A i n_B niech będą liczbami cząstek zaobserwowanych tylko przez A i tylko przez B , oraz niech N_C będzie liczbą cząstek zaobserwowanych przez oba detektory jednocześnie. Liczby cząstek zaobserwowanych przez każdy z detektorów to $N_B = n_B + N_C$ oraz $N_A = n_A + N_C$.
- Każdą cząstkę przechodzącą przez oba detektory możemy zakwalifikować do jednej z czterech rozłącznych klas:
 - cząstka zarejestrowana tylko przez A : $p_a = p_A(1 - p_B)$, zmienna losowa $n_A = N_A - N_C$
 - cząstka zarejestrowana tylko przez B : $p_b = p_B(1 - p_A)$, zmienna losowa $n_B = N_B - N_C$
 - cząstka zarejestrowana przez A i B : $p_C = p_A p_B$, zmienna losowa N_C
 - brak rejestracji przez A i B : $p = (1 - p_A)(1 - p_B)$, $n = N - n_A - n_B - N_C$
- Rozkład p -twa zmiennych (n_A, n_B, N_C, N) zadany jest przez:

$$P_{n_A, n_B, N_C, N}(p_A, p_B, \mu) = W_{n_A, n_B, N_C}(N, p_A, p_B) \mathcal{P}_N(\mu)$$

gdzie

$$W_{n_A, n_B, N_C}(N, p_A, p_B) = \frac{N!}{n_A! n_B! N_C! n!} p_a^{n_A} p_b^{n_B} p_C^{N_C} p^n$$

Efektywność detektora

- Ponieważ:
$$\prod_{i=1}^j \mathcal{P}_{n_i}(\mu_i) = \frac{\mu_1^{n_1}}{n_1!} e^{-\mu_1} \dots \frac{\mu_j^{n_j}}{n_j!} e^{-\mu_j} = \frac{\mu^n}{n!} e^{-\mu} p_1^{n_1} p_2^{n_2} \dots p_j^{n_j} =$$
$$= \mathcal{P}_n(\mu) \mathcal{W}_{n_1 \dots n_j}(n, p_1, \dots, p_j)$$

gdzie
$$n = \sum_{i=1}^j n_i, \quad \mu = \sum_{i=1}^j \mu_i, \quad p_i = \frac{\mu_i}{\mu}$$

więc łączny rozkład p-twa przyjmuje postać:

$$\mathcal{P}_{n_A, n_B, N_C, n}(p_A, p_B, \mu) = \mathcal{P}_{n_A}(\mu p_A) \mathcal{P}_{n_B}(\mu p_B) \mathcal{P}_{N_C}(\mu p_C) \mathcal{P}_n(\mu p)$$

- Po wysumowaniu po n , możemy zapisać funkcję wiarygodności:

$$\mathcal{L}(n_A, n_B, N_C; p_A, p_B, \mu) = \mathcal{P}_{n_A}(\mu p_A) \mathcal{P}_{n_B}(\mu p_B) \mathcal{P}_{N_C}(\mu p_C)$$

- Obliczamy $\ln \mathcal{L}$, różniczkujemy po parametrach i znajdujemy estymatory:

$$\left. \begin{aligned} \frac{\partial \ln \mathcal{L}}{\partial p_A} &= \frac{n_A + N_C}{p_A} - \frac{n_B}{1 - p_A} - \mu(1 - p_B) = 0 \\ \frac{\partial \ln \mathcal{L}}{\partial p_B} &= \frac{n_B + N_C}{p_B} - \frac{n_A}{1 - p_B} - \mu(1 - p_A) = 0 \\ \frac{\partial \ln \mathcal{L}}{\partial \mu} &= \frac{n_A + n_B + N_C}{\mu} - (1 - (1 - p_A)(1 - p_B)) = 0 \end{aligned} \right\} \Rightarrow \left\{ \begin{aligned} \hat{p}_A &= \frac{N_C}{N_B} \\ \hat{p}_B &= \frac{N_C}{N_A} \\ \hat{\mu} &= \frac{N_A N_B}{N_C} \end{aligned} \right.$$

- Korzystając z tw. Cramera-Rao znajdujemy macierz kowariancji $V(\hat{p}_A, \hat{p}_B, \hat{\mu})$:

$$V = \begin{bmatrix} \frac{p_A(1-p_A)}{\mu p_B} & \frac{(1-p_A)(1-p_B)}{\mu} & -\frac{(1-p_A)(1-p_B)}{p_B} \\ \frac{(1-p_A)(1-p_B)}{\mu} & \frac{p_B(1-p_B)}{\mu p_A} & -\frac{(1-p_A)(1-p_B)}{p_A} \\ -\frac{(1-p_A)(1-p_B)}{p_B} & -\frac{(1-p_A)(1-p_B)}{p_A} & \mu \frac{(1-p_A)(1-p_B) + p_A p_B}{p_A p_B} \end{bmatrix}$$

- Po wstawieniu za parametry ich estymatorów otrzymujemy $\hat{V}(\hat{p}_A, \hat{p}_B, \hat{\mu})$
- Ze względu na "regułę zatrzymania" (Poissona), nawet gdy wydajność jednego z detektorów wynosi 100%, to $\mathcal{V}[N] \neq 0$.
- Rozważmy regułę zatrzymania w postaci zadanej liczby $n_{AB} = n_A + n_B + N_C$ cząstek łącznie zarejestrowanych przez oba detektory.
- Łączny rozkład p-twa zmiennych n_A oraz n_B przy ustalonej wartości n_{AB} :

$$\begin{aligned} P_{n_A, n_B}(n_{AB}, p_A, p_B) &= \frac{\mathcal{W}_{n_A, n_B, N_C}(N, p_A, p_B)}{\mathcal{B}_{n_{AB}}(N, p_{AB})} = \\ &= \frac{n_{AB}!}{n_A! n_B! (N_{AB} - n_A - n_B)!} \left(\frac{p_a}{p_{AB}}\right)^{n_A} \left(\frac{p_b}{p_{AB}}\right)^{n_B} \left(\frac{p_C}{p_{AB}}\right)^{n_{AB} - n_A - n_B} \end{aligned}$$

- Rozkład $P_{n_A, n_B}(n_{AB}, p_A, p_B)$ odgrywa rolę funkcji wiarygodności, więc:

$$\ln \mathcal{L} = N_A \ln p_A + n_B \ln(1 - p_A) + N_B \ln p_B + n_A \ln(1 - p_B) - n_{AB} \ln(p_A + p_B - p_{APB}) + \text{const}$$

- Korzystając z tw. Cramera-Rao znajdujemy macierz kowariancji $V(\hat{p}_A, \hat{p}_B)$:

$$V = \begin{bmatrix} \frac{p_A(1-p_A)}{N_B} & \frac{(1-p_A)(1-p_B)(p_A+p_B-p_{APB})}{n_{AB}} \\ \frac{(1-p_A)(1-p_B)(p_A+p_B-p_{APB})}{n_{AB}} & \frac{p_B(1-p_B)}{N_A} \end{bmatrix}$$

- Całkowita liczba cząstek jest zmienną losową podlegającą rozkładowi ujemnemu dwumianowemu, gdzie liczba sukcesów to n_{AB} , a p -two sukcesu to p_{AB} .

Za ocenę nieznannej liczby cząstek N i jej wariancję, możemy przyjąć odpowiednio wartość oczekiwaną i wariancję dla zmiennej z tego rozkładu:

$$\mathcal{E}[N] = \frac{n_{AB}}{p_{AB}} \quad \mathcal{V}[N] = n_{AB} \frac{1-p_{AB}}{p_{AB}^2} = \frac{n_{AB}}{p_{AB}} \frac{(1-p_A)(1-p_B)}{1-(1-p_A)(1-p_B)}$$

- Widać, że jeśli jedna z efektywności dąży to jedności, to wariancja dąży do zera.

Zastosowanie MNW do danych zgrupowanych

- Rozważamy histogram o n przedziałach $[x_k - \Delta_k/2, x_k + \Delta_k/2]$, $k = 1, \dots, n$ i liczbie przypadków w przedziale n_k oraz całkowitej liczbie przypadków N .
- P-two znalezienia zmiennej losowej x w przedziale o środku x_k :

$$P_k(\vec{\theta}) = \int_{x_k - \Delta_k/2}^{x_k + \Delta_k/2} f(x; \vec{\theta}) dx, \quad \sum_{k=1}^n P_k(\vec{\theta}) = 1$$

- Jeśli liczba przypadków N jest ustalona (dośw. Bernoulliego), to liczby n_k pochodzą z rozkładu wielomianowego i funkcja wiarygodności ma postać:

$$\mathcal{L}(\vec{\theta}) \equiv \mathcal{W}_{n_1, \dots, n_n}(N, P_1, \dots, P_n) = \frac{N!}{n_1! n_2! \dots n_n!} P_1^{n_1} P_2^{n_2} \dots P_n^{n_n}$$

- Jeśli liczba przypadków N podlega rozkładowi Poissona o parametrze μ , to funkcja wiarygodności przyjmuje postać ($\mu_k = \mu P_k(\vec{\theta})$, $\mu = \sum_{k=1}^n \mu_k$):

$$\mathcal{L}(\vec{\theta}) = \prod_{k=1}^n \mathcal{P}_{n_k}(\mu_k) = \prod_{k=1}^n \frac{\mu_k^{n_k}}{n_k!} e^{-\mu_k} = \frac{\mu^N}{N!} e^{-\mu} \frac{N!}{n_1! \dots n_n!} P_1^{n_1} \dots P_n^{n_n}$$

- Zakładając, że $\mathcal{V}[n_k] = \mu P_k(1 - P_k) \approx \mu P_k = n_k$, w obu przypadkach maksymalizujemy wyrażenie:

$$\ln \mathcal{L}(\vec{\theta}) = \sum_{k=1}^n n_k \ln P_k(\vec{\theta}) + \text{const} \approx -\frac{1}{2} \sum_{k=1}^n \left(\frac{n_k - N P_k(\vec{\theta})}{\sqrt{n_k}} \right)^2$$

Metoda Najmniejszych Kwadratów (MNK)

- Załóżmy, że mierzymy n niezależnych wartości, y_1, y_2, \dots, y_n , każda z rozkładu Gaussa o nieznannej wartości oczekiwanej i znanej wariancji:

$$\mathcal{E}[y_i] = \eta(x_i; \vec{\theta}) \quad \mathcal{V}[y_i] = \sigma_i^2$$

O tzw. zmiennych kontrolowanych x_i zakładamy, że są znane dokładnie.

- Chcemy oszacować nieznanne parametry $\vec{\theta}$ tak aby dopasowanie krzywej do punktów pomiarowych było możliwie najlepsze.
- Konstruujemy łączną funkcję gęstości p-twa:

$$g(y_1, \dots, y_n; \eta_1, \dots, \eta_n, \sigma_1, \dots, \sigma_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - \eta_i)^2}{2\sigma_i^2}\right)$$

Wówczas log z funkcji wiarygodności ma postać (pomijamy wyrazy niezależne od parametrów):

$$\ln \mathcal{L}(\vec{\theta}) = -\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \eta(x_i; \vec{\theta}))^2}{\sigma_i^2} + \dots$$

- Zasada najmniejszych kwadratów głosi, że parametry należy dobrać tak, aby spełniony był warunek:

$$\mathcal{R}(\vec{\theta}) = \sum_{i=1}^n \frac{(y_i - \eta(x_i; \vec{\theta}))^2}{\sigma_i^2} = \min(\vec{\theta})$$

- Funkcję $\eta(x_i; \hat{\theta})$ nazywamy krzywą regresji najlepszego dopasowania MNK. Parametry $\hat{\theta}$ to estymatory MNK.
- W przypadku skorelowanych pomiarów, musimy skorzystać z wielowymiarowego rozkładu Gaussa ze znaną macierzą kowariancji V .
- Wówczas MNK sprowadza się do warunku:

$$\mathcal{R}(\vec{\theta}) = \sum_{i,j=1}^n \left(y_i - \eta(x_i; \vec{\theta}) \right) (V^{-1})_{ij} \left(y_j - \eta(x_j; \vec{\theta}) \right) = \min(\vec{\theta})$$

MNK - przypadek liniowy

- Rozważmy sytuację kiedy związek pomiędzy parametrami θ_i , a wielkością mierzoną η przyjmuje postać liniową:

$$\eta(x; \vec{\theta}) = \varphi_1(x)\theta_1 + \varphi_2(x)\theta_2 + \varphi_m(x)\theta_m$$

gdzie $\varphi_i(x)$ są znanymi, liniowo niezależnymi funkcjami.

- Dla n punktów pomiarowych (x_i, y_i) , $i = 1, \dots, n$ otrzymujemy układ n równań:

$$\begin{cases} \eta_1 = \varphi_1(x_1)\theta_1 + \varphi_2(x_1)\theta_2 + \varphi_m(x_1)\theta_m \\ \eta_2 = \varphi_1(x_2)\theta_1 + \varphi_2(x_2)\theta_2 + \varphi_m(x_2)\theta_m \\ \vdots \\ \eta_n = \varphi_1(x_n)\theta_1 + \varphi_2(x_n)\theta_2 + \varphi_m(x_n)\theta_m \end{cases} \Leftrightarrow \vec{\eta} = \Phi \vec{\theta}$$

gdzie

$$\vec{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix} \quad \Phi = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \dots & \varphi_m(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \dots & \varphi_m(x_2) \\ \vdots & & \ddots & \vdots \\ \varphi_1(x_n) & \varphi_2(x_n) & \dots & \varphi_m(x_n) \end{pmatrix} \quad \vec{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{pmatrix}$$

Macierz Φ nazywamy macierzą konstrukcyjną.

- Zasada najmniejszych kwadratów przyjmuje postać (oznaczenie $Q = V^{-1}$):

$$\mathcal{R} = (\vec{y} - \Phi\vec{\theta})^T Q (\vec{y} - \Phi\vec{\theta}) = \min(\vec{\theta})$$

Zapiszemy wielkość \mathcal{R} w jawnej postaci:

$$\begin{aligned} \mathcal{R} &= \sum_{i,j=1}^n \left(y_i - \sum_{k=1}^m \varphi_k(x_i) \theta_k \right) Q_{ij} \left(y_j - \sum_{l=1}^m \varphi_l(x_j) \theta_l \right) = \\ &= \sum_{i,j=1}^n y_i Q_{ij} y_j - 2 \sum_{i,j=1}^n \sum_{k=1}^m \varphi_k(x_i) \theta_k Q_{ij} y_j + \sum_{i,j=1}^n \sum_{k,l=1}^m \varphi_k(x_i) \theta_k Q_{ij} \varphi_l(x_j) \theta_l \end{aligned}$$

Różniczkujemy względem parametru θ_p ($p = 1, \dots, m$):

$$\begin{aligned} \frac{\partial \mathcal{R}}{\partial \theta_p} &= -2 \sum_{i,j=1}^n \sum_{k=1}^m \varphi_k(x_i) \delta_{kp} Q_{ij} y_j + \sum_{i,j=1}^n \sum_{k,l=1}^m \varphi_k(x_i) \delta_{kp} Q_{ij} \varphi_l(x_j) \theta_l + \\ &\quad + \sum_{i,j=1}^n \sum_{k,l=1}^m \varphi_k(x_i) \theta_k Q_{ij} \varphi_l(x_j) \delta_{lp} = \\ &= -2 \sum_{i,j=1}^n \varphi_p(x_i) Q_{ij} y_j + \sum_{i,j=1}^n \sum_{l=1}^m \varphi_p(x_i) Q_{ij} \varphi_l(x_j) \theta_l + \\ &\quad + \sum_{i,j=1}^n \sum_{k=1}^m \varphi_k(x_i) \theta_k Q_{ij} \varphi_p(x_j) \end{aligned}$$

- Zamieniając indeks l na k oraz zmieniając pomiędzy sobą nazwy indeksów i oraz j otrzymujemy (macierz $Q = V^{-1}$ jest symetryczna):

$$\frac{\partial \mathcal{R}}{\partial \theta_p} = -2 \underbrace{\sum_{i,j=1}^n \varphi_p(x_i) Q_{ij} y_j}_{(\Phi^T Q \vec{y})_p} + 2 \underbrace{\sum_{i,j=1}^n \sum_{l=1}^m \varphi_p(x_i) Q_{ij} \varphi_l(x_j) \theta_l}_{(\Phi^T Q \Phi \vec{\theta})_p} = 0$$

- Otrzymaliśmy układ liniowych równań (zwanymi normalnymi) na nieznanne parametry θ_i w postaci macierzowej:

$$\Phi^T Q \Phi \vec{\theta} = \Phi^T Q \vec{y}$$

o rozwiązaniach liniowo zależnych od mierzonych wielkości:

$$\hat{\vec{\theta}} = (\Phi^T Q \Phi)^{-1} \Phi^T Q \vec{y} = \underbrace{(\Phi^T V^{-1} \Phi)^{-1}}_{\equiv W} \Phi^T V^{-1} \vec{y} = \underbrace{W \Phi^T V^{-1}}_{\Psi} \vec{y} = \Psi \vec{y}$$

- Jeśli wielkości mierzone są nieobciążone, to również nieobciążone są estymatory parametrów:

$$\mathcal{E}[\hat{\vec{\theta}}] = \mathcal{E}[\Psi \vec{y}] = \Psi \mathcal{E}[\vec{y}] = \Psi \vec{\eta} = (\Phi^T V^{-1} \Phi)^{-1} \Phi^T V^{-1} \Phi \vec{\theta} = \vec{\theta}$$

- Macierz kowariancji estymatorów parametrów ma postać:

$$\begin{aligned} V[\hat{\theta}] &= \mathcal{E}[\Psi(\vec{y} - \vec{\eta})(\Psi(\vec{y} - \vec{\eta}))^T] = \mathcal{E}[\Psi(\vec{y} - \vec{\eta})(\vec{y} - \vec{\eta})^T \Psi^T] = \\ &= \Psi \mathcal{E}[(\vec{y} - \vec{\eta})(\vec{y} - \vec{\eta})^T] \Psi^T = W \Phi^T V^{-1} \mathcal{E}[(\vec{y} - \vec{\eta})(\vec{y} - \vec{\eta})^T] V^{-1} \Phi W = \\ &= W \Phi^T V^{-1} V V^{-1} \Phi W = W \underbrace{\Phi^T V^{-1} \Phi}_{W^{-1}} W = W \end{aligned}$$

A więc: $V[\hat{\theta}] = W = (\Phi^T V^{-1} \Phi)^{-1}$

- Dysponując estymatami parametrów możemy wykorzystać je do interpolacji bądź ekstrapolacji, konstruując krzywą najlepszego dopasowania:

$$\hat{\eta}(x) = \varphi_1(x)\hat{\theta}_1 + \varphi_2(x)\hat{\theta}_2 + \dots + \varphi_m(x)\hat{\theta}_m = \vec{\varphi}^T(x)\hat{\theta}$$

Błąd takiej operacji wynosi:

$$\begin{aligned} \mathcal{V}[\hat{\eta}(x)] &= \mathcal{E}[(\hat{\eta}(x) - \eta(x))^2] = \mathcal{E}[(\vec{\varphi}^T(x)(\hat{\theta} - \vec{\theta}))(\vec{\varphi}^T(x)(\hat{\theta} - \vec{\theta}))^T] = \\ &= \mathcal{E}[\vec{\varphi}^T(x)(\hat{\theta} - \vec{\theta})(\hat{\theta} - \vec{\theta})^T \vec{\varphi}(x)] = \vec{\varphi}^T(x) W \vec{\varphi}(x) \end{aligned}$$

- **Twierdzenie Gaussa-Markowa:** Pośród wszystkich nieobciążonych estymatorów, które są liniowymi kombinacjami wielkości mierzonych, estymatory metody najmniejszych kwadratów mają najmniejszą wariancję.

Ocena jakości dopasowania

- Ocena jakości dopasowania na podstawie reszkowej różnicy w minimum:

$$\vec{\varepsilon}_{\min} = \vec{y} - \hat{\eta}$$

- Jej wartość oczekiwana i wariancja są odpowiednio równe:

$$\mathcal{E}[\vec{\varepsilon}_{\min}] = \mathcal{E}[\vec{y} - \hat{\eta}] = \mathcal{E}[\vec{y}] - \mathcal{E}[\hat{\eta}] = \vec{\eta} - \mathcal{E}[\Psi\hat{\theta}] = \vec{\eta} - \Psi\vec{\theta} = \vec{\eta} - \vec{\eta} = 0$$

$$\mathcal{V}[\vec{\varepsilon}_{\min}] = \mathcal{E}[(\vec{y} - \hat{\eta})(\vec{y} - \hat{\eta})^T] = \dots = \mathcal{V}[\vec{y}] - \mathcal{V}[\hat{\eta}]$$

- Ocena błędów systematycznych – wpływ (ang. pull): $z_i = \frac{y_i - \hat{\eta}_i}{\sqrt{\mathcal{V}[y_i] - \mathcal{V}[\hat{\eta}_i]}}$

Wartość średnia wpływu istotnie różna od 0 wskazuje na błędy systematyczne w danych, a odchylenie standardowe istotnie różne od 1 na niedostateczną kontrolę nad błędami statystycznymi procedury pomiarowej.

- Ocena jakości dopasowania na podstawie wartości \mathcal{R}_{\min} :
 - porównanie modeli – im mniejsze \mathcal{R}_{\min} tym lepsza zgodność (uwaga: dopasowując wielomian $n - 1$ stopnia do n punktów dostaniemy $\mathcal{R}_{\min} = 0$).
 - gdy mierzone wielkości pochodzą z rozkładu normalnego, wówczas estymatory parametrów mają rozkłady normalne, natomiast wielkość \mathcal{R}_{\min} jest statystyką χ^2 wylosowaną z rozkładu o $(n - m)$ stopniach swobody $\mathcal{R}_{\min} = \chi_{n-m}^2$ o ile postulowana zależność $f(x; \theta_1, \dots, \theta_m)$ jest słuszna. (Jeśli macierz $V(y)$ jest macierzą kowariancji, a nie jej estymat.)