

Analiza danych

Mariusz Przybycień

Wydział Fizyki i Informatyki Stosowanej
Akademia Górniczo-Hutnicza

Wykład 6

- Rozważmy dopasowanie linii prostej do danych doświadczalnych:

$$\vec{\eta} = \Phi \vec{\theta} \quad \Leftrightarrow \quad \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

- Zakładamy, że pomiary nie są skorelowane:

$$V = \begin{bmatrix} s_1^2 & 0 & \dots & 0 \\ 0 & s_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_n^2 \end{bmatrix} \Rightarrow V^{-1} = \begin{bmatrix} s_1^{-2} & 0 & \dots & 0 \\ 0 & s_2^{-2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_n^{-2} \end{bmatrix}$$

- Wprowadzamy oznaczenia:

$$S = \sum_{i=1}^n \frac{1}{s_i^2} \quad S_x = \sum_{i=1}^n \frac{x_i}{s_i^2} \quad S_{xx} = \sum_{i=1}^n \frac{x_i^2}{s_i^2} \quad S_{xy} = \sum_{i=1}^n \frac{x_i y_i}{s_i^2}$$
$$S_y = \sum_{i=1}^n \frac{y_i}{s_i^2} \quad S_{yy} = \sum_{i=1}^n \frac{y_i^2}{s_i^2} \quad \Delta = \frac{1}{SS_{xx} - S_x^2}$$

- Konstruujemy macierze W i Ψ :

$$\begin{aligned} \Phi^T V^{-1} \Phi &= \\ &= \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} s_1^{-2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & s_n^{-2} \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} \frac{1}{s_1^2} & \dots & \frac{1}{s_n^2} \\ x_1 & \dots & x_n \\ \frac{1}{s_1^2} & \dots & \frac{1}{s_n^2} \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} S & S_x \\ S_x & S_{xx} \end{bmatrix} \end{aligned}$$

$$W = (\Phi^T V^{-1} \Phi)^{-1} = \Delta \begin{bmatrix} S_{xx} & -S_x \\ -S_x & S \end{bmatrix}$$

$$\begin{aligned} \Psi &= W \Phi^T V^{-1} = \\ &= \Delta \begin{bmatrix} S_{xx} & -S_x \\ -S_x & S \end{bmatrix} \begin{bmatrix} \frac{1}{s_1^2} & \dots & \frac{1}{s_n^2} \\ x_1 & \dots & x_n \\ \frac{1}{s_1^2} & \dots & \frac{1}{s_n^2} \end{bmatrix} = \Delta \begin{bmatrix} \frac{S_{xx} - x_1 S_x}{s_1^2} & \dots & \frac{S_{xx} - x_n S_x}{s_n^2} \\ x_1 S - S_x & \dots & x_n S - S_x \\ \frac{1}{s_1^2} & \dots & \frac{1}{s_n^2} \end{bmatrix} \end{aligned}$$

- Wyznaczamy estymatory parametrów linii prostej:

$$\hat{\theta} = \Psi \vec{y} \Rightarrow \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \Delta \begin{bmatrix} S_{xx} S_y - S_{xy} S_x \\ S S_{xy} - S_x S_y \end{bmatrix}$$

- Z macierzy $W = V[\hat{\theta}]$ odczytujemy:

$$\begin{aligned} \mathcal{V}[\hat{\theta}_1] &= S_{xx}\Delta & \text{cov}[\hat{\theta}_1, \hat{\theta}_2] &= -S_x\Delta & \rho(\hat{\theta}_1, \hat{\theta}_2) &= -\frac{S_x}{\sqrt{SS_{xx}}} \\ \mathcal{V}[\hat{\theta}_2] &= S\Delta \end{aligned}$$

- Obliczamy wartość statystyki \mathcal{R} w minimum: $\mathcal{R}_{\min} = S_{yy} - S_y\hat{\theta}_1 - S_{xy}\hat{\theta}_2$
- Ponieważ parametry są liniowymi funkcjami wielkości mierzonych, więc poziomicę stałej wartości \mathcal{R} (czyli χ^2) są elipsami:

$$\frac{1}{1 - \rho^2} \left[\frac{(\theta_1 - \hat{\theta}_1)^2}{\mathcal{D}^2[\hat{\theta}_1]} + \frac{(\theta_2 - \hat{\theta}_2)^2}{\mathcal{D}^2[\hat{\theta}_2]} - 2\rho \left(\frac{\theta_1 - \hat{\theta}_1}{\mathcal{D}[\hat{\theta}_1]} \right) \left(\frac{\theta_2 - \hat{\theta}_2}{\mathcal{D}[\hat{\theta}_2]} \right) \right] = \mathcal{R} - \mathcal{R}_{\min}$$

- Problem dopasowania do danych wielu nieliniowych funkcji można sprowadzić do problemu dopasowania linii prostej:

$$y = ab^x$$

$$y' = \ln a + x \ln b = a' + b'x$$

$$y = ax^b$$

$$y' = \ln a + b \ln x = a' + bx'$$

$$y = a \exp [bx]$$

$$y' = \ln a + bx = a' + bx$$

$$y = \frac{x}{a + bx}$$

$$y' = a\frac{1}{x} + b = b + ax'$$

Dopasowanie linii prostej - dygresja

- Dopasowanie linii prostej z pominięciem rachunku macierzowego:

$$\eta(x) = f(x; \vec{\theta}) = \theta_1 + \theta_2 x$$

- Znajdujemy wartości parametrów dla których \mathcal{R} osiąga minimum:

$$\begin{cases} \frac{\partial \mathcal{R}}{\partial \theta_1} = -2 \sum_{i=1}^n \left(\frac{y_i - \theta_1 - \theta_2 x_i}{s_i^2} \right) = -2 \left(\sum_{i=1}^n \frac{y_i}{s_i^2} - \theta_1 \sum_{i=1}^n \frac{1}{s_i^2} - \theta_2 \sum_{i=1}^n \frac{x_i}{s_i^2} \right) \\ \frac{\partial \mathcal{R}}{\partial \theta_2} = -2 \sum_{i=1}^n x_i \left(\frac{y_i - \theta_1 - \theta_2 x_i}{s_i^2} \right) = -2 \left(\sum_{i=1}^n \frac{x_i y_i}{s_i^2} - \theta_1 \sum_{i=1}^n \frac{x_i}{s_i^2} - \theta_2 \sum_{i=1}^n \frac{x_i^2}{s_i^2} \right) \end{cases}$$

- Przyrównując oba równania do zera otrzymujemy:

$$\hat{\theta}_1 = \Delta(S_{xx}S_y - S_x S_{xy}) \quad \hat{\theta}_2 = \Delta(SS_{xy} - S_x S_y)$$

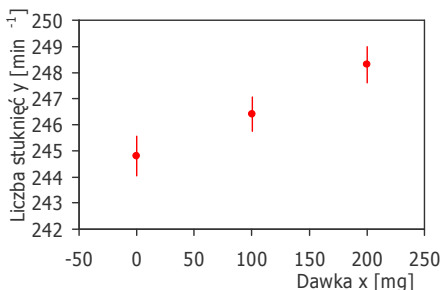
- Prosta najlepszego dopasowania, $\hat{\eta}(x) = \hat{\theta}_1 + \hat{\theta}_2 x$, pozwala dokonać interpolacji lub ekstrapolacji. Błąd tej operacji dany jest przez:

$$\begin{aligned} \mathcal{V}[\hat{\eta}] &= \varphi(x) W \varphi^T(x) = \Delta [1 \ x] \begin{bmatrix} S_{xx} & -S_x \\ -S_x & S \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \Delta(S_{xx} - 2xS_x + x^2S) = \\ &= \Delta S \left(\frac{SS_{xx} - S_x^2}{S^2} + \left(x - \frac{S_x}{S} \right)^2 \right) = \frac{1}{S} + \mathcal{V}[\hat{\theta}_2] \left(x - \frac{S_x}{S} \right)^2 \end{aligned}$$

MNK - dopasowanie linii prostej

Przykład: Zbadano wpływ kofeiny na sprawność wykonywania prostych czynności manualnych. 30 studentów podzielono losowo na trzy grupy po 10 osób i każdej grupie podano różne ilości kofeiny: 0 mg, 100 mg i 200 mg. Po dwóch godzinach poroszono ich o wykonanie z maksymalną szybkością czynności polegającej na stukaniu palcami po stole. Wyniki pomiarów przedstawia tabela:

Dawka x [mg]	Liczba stuknięć palcami y [min^{-1}]										$\bar{y} \pm \sigma_{\bar{y}}$
0	242	245	244	248	247	248	242	244	246	242	244.8 ± 0.76
100	248	246	245	247	248	250	247	246	243	244	246.4 ± 0.65
200	246	248	250	252	248	250	246	248	245	250	248.3 ± 0.70



Prosta dopasowana MNK:

$$y = \underbrace{244.7 \pm 0.7}_{\hat{\theta}_1 \pm \hat{\sigma}_{\theta_1}} + \underbrace{0.017 \pm 0.005}_{\hat{\theta}_2 \pm \hat{\sigma}_{\theta_2}} x$$

Kowariancja i współczynnik korelacji pomiędzy estymatorami parametrów:

$$\text{cov}[\hat{\theta}_1, \hat{\theta}_2] = -0.0028$$

$$\rho(\hat{\theta}_1, \hat{\theta}_2) = -0.8$$

MNK - dopasowanie linii prostej

Niepewność inter- i ekstrapolacji:

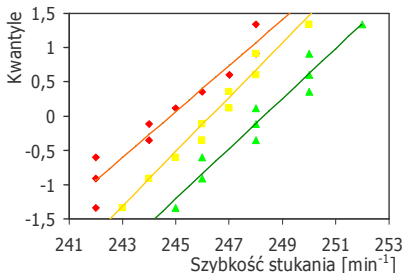
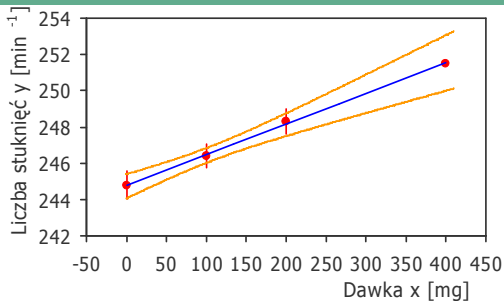
$$S = 6.139 \quad S_x = 644.85$$

$$\hat{\eta}(x) =$$

$$\hat{\theta}_1 + \hat{\theta}_2 x \pm \sqrt{\frac{1}{S} + \mathcal{V}[\hat{\theta}_2] \left(x - \frac{S_x}{S}\right)^2}$$

Przewidywanie dla $x = 400$:

$$\eta(400) = 251.5 \pm 1.5$$



Dystrybuanta empiryczna:

$$\hat{F}(x) = \begin{cases} 0 & x < x_1 \\ \frac{i}{n+1} & x_i \leq x < x_{i+1} \\ \frac{n}{n+1} & x \geq x_n \end{cases}$$

$$F(q_i) = \frac{i}{n+1} = \frac{1}{11} \quad q = \frac{x - \mu}{\sigma}$$

$F(q_i)$	0.091	0.182	0.273	0.364	0.455	0.546	0.636	0.727	0.818	0.909
q_i	-1.335	-0.908	-0.605	-0.349	-0.114	0.114	0.349	0.605	0.908	1.335

- Jeśli hipoteza o zależności liniowej jest prawdziwa, to statystyka \mathcal{R} ma rozkład χ^2 o $n - m = 3 - 2 = 1$ stopniach swobody:

$$\mathcal{R}_{\min} = S_{yy} - S_y \hat{\theta}_1 - S_{xy} \hat{\theta}_2 = 0.033 \Rightarrow P(\chi^2 > \mathcal{R}_{\min}) = \int_{\mathcal{R}_{\min}}^{\infty} \chi_1(u) du = 0.857$$

- Jeśli zażądamy $P(\chi^2 > \mathcal{R}_{\min}) = \int_{\mathcal{R}_{\min}}^{\infty} \chi_1(u) du = 0.99 \Rightarrow \mathcal{R}_{\min} = 0.00015$

- Oznacza to, że średnie odchylenie każdego punktu od dopasowanej linii w jednostkach typowej niepewności wynosi tylko: $\sqrt{0.00015/3} \approx 0.007$

- Ponieważ parametry są liniowymi funkcjami wielkości mierzonych, więc poziomice stałej wartości \mathcal{R} (czyli χ^2) są elipsami:

$$\frac{1}{1 - \rho^2} \left[\frac{(\theta_1 - \hat{\theta}_1)^2}{D^2[\hat{\theta}_1]} + \frac{(\theta_2 - \hat{\theta}_2)^2}{D^2[\hat{\theta}_2]} - 2\rho \left(\frac{\theta_1 - \hat{\theta}_1}{D[\hat{\theta}_1]} \right) \left(\frac{\theta_2 - \hat{\theta}_2}{D[\hat{\theta}_2]} \right) \right] = \mathcal{R} - \mathcal{R}_{\min}$$

- Jeśli nie jest znana macierz kowariancji wielkości mierzonych, to MNK pozwala znaleźć jej estymatę w pewnych szczególnych sytuacjach, np. gdy macierz ta znana jest z dokładnością do stałego czynnika, $V = \sigma^2 \tilde{V}$:

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

$$V[\hat{\theta}] = W = (\Phi^T V^{-1} \Phi)^{-1} = \sigma^2 (\Phi^T \tilde{V}^{-1} \Phi)^{-1} \equiv \sigma^2 \tilde{W}$$

- Minimalna ważona suma kwadratów reszt, $\hat{\varepsilon}_i = y_i - \hat{\eta}_i = y_i - f(x_i, \hat{\theta})$, przyjmuje postać:

$$\mathcal{R}_{\min} = \hat{\varepsilon}^T V^{-1} \hat{\varepsilon} = (\vec{y}^T - \hat{\theta}^T \Phi^T) V^{-1} (\vec{y} - \Phi \hat{\theta}) =$$

$$\left\{ \Phi^T V^{-1} \Phi \vec{\theta} = \Phi^T V^{-1} \vec{y} \Rightarrow \Phi^T V^{-1} (\Phi \vec{\theta} - \vec{y}) = 0 \Rightarrow (\Phi \vec{\theta} - \vec{y})^T V^{-1} \Phi = 0 \right\}$$

$$= (\vec{y}^T - \hat{\theta}^T \Phi^T) V^{-1} \vec{y} = \vec{y}^T V^{-1} \vec{y} - \hat{\theta}^T \Phi^T V^{-1} \vec{y} = \vec{y}^T V^{-1} \vec{y} - \hat{\theta}^T W^{-1} \hat{\theta} =$$

$$= \left\{ \vec{\eta} = \Phi \vec{\theta} \right\} = (\vec{y} - \vec{\eta})^T V^{-1} (\vec{y} - \vec{\eta}) - (\hat{\theta} - \vec{\theta})^T W^{-1} (\hat{\theta} - \vec{\theta})$$

- Można pokazać, że $\mathcal{E}[\mathcal{R}_{\min}] = \text{Tr}(V^{-1}V) - \text{Tr}(W^{-1}W) = n - m$

- Można więc zaproponować nieobciążony i niezależny od rozkładu estymator $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \tilde{V}^{-1} \hat{\varepsilon}}{n - m}$$

- W konsekwencji otrzymujemy pełną postać estymatorów macierzy kowariancji mierzonych wielkości oraz estymowanych parametrów:

$$\hat{V} = \hat{\sigma}^2 \tilde{V}, \quad \hat{W} = \hat{\sigma}^2 \left(\Phi^T \tilde{V}^{-1} \Phi \right)^{-1} \equiv \hat{\sigma}^2 \tilde{W}$$

- Przykład: Rozważmy dopasowanie prostej $\eta(x) = \theta_1 + \theta_2 x$ do nieskorelowanych danych (x_i, y_i) , których błędy znane są z dokładnością do czynnika skalującego, $s_i = \sigma u_i$:

$$S = \frac{1}{\sigma^2} \sum_{i=1}^n \frac{1}{u_i^2} \equiv \frac{1}{\sigma^2} U, \quad S_x = \frac{1}{\sigma^2} \sum_{i=1}^n \frac{x_i}{u_i^2} \equiv \frac{1}{\sigma^2} U_x, \quad S_y = \frac{1}{\sigma^2} \sum_{i=1}^n \frac{y_i}{u_i^2} \equiv \frac{1}{\sigma^2} U_y$$

$$S_{xx} = \frac{1}{\sigma^2} \sum_{i=1}^n \frac{x_i^2}{u_i^2} \equiv \frac{1}{\sigma^2} U_{xx}, \quad S_{yy} = \frac{1}{\sigma^2} \sum_{i=1}^n \frac{y_i^2}{u_i^2} \equiv \frac{1}{\sigma^2} U_{yy}$$

$$\Delta = \frac{1}{S S_{xx} - S_x^2} = \frac{\sigma^4}{U U_{xx} - U_x^2} \equiv \sigma^4 \delta$$

Estymatory parametrów linii prostej oraz ich macierzy kowariancji:

$$\hat{\theta}_1 = (U_{xx}U_y - U_xU_{xy})\delta, \quad \hat{\theta}_2 = (UU_{xy} - U_xU_y)\delta$$

$$\hat{\sigma}^2 = \frac{(\vec{y} - \hat{\theta}_1 - \hat{\theta}_2\vec{x})^T \tilde{V}^{-1} (\vec{y} - \hat{\theta}_1 - \hat{\theta}_2\vec{x})}{n-2}$$

$$\hat{V} [\hat{\theta}_1, \hat{\theta}_2] = \begin{bmatrix} U_{xx}\hat{\sigma}^2\delta & -U_x\hat{\sigma}^2\delta \\ -U_x\hat{\sigma}^2\delta & U\hat{\sigma}^2\delta \end{bmatrix}$$

W zależności od stanu wiedzy o błędach s_i należy odpowiednio wybrać u_i :

- stały błąd bezwzględny: $s_i = \sigma \Rightarrow u_i = 1$
- stały błąd względny: $s_i = \sigma y_i \Rightarrow u_i = y_i$
- Rozważmy zmianę zmiennych: $y = \frac{x}{ax+b}, z = \frac{1}{y} = a\frac{1}{x} + b = ax' + b \Rightarrow \sigma_z = \frac{\sigma_y}{y^2}$

stały błąd bezwzględny: $\sigma_z = s_i \frac{1}{y_i^2} \Rightarrow u_i = \frac{1}{y_i^2}$

stały błąd względny: $\sigma_z = \frac{s_i}{y_i} \frac{1}{y_i} \Rightarrow u_i = \frac{1}{y_i}$

- Rozważmy przypadek skorelowanych pomiarów tej samej wielkości, $\eta(x) = \theta$.

$$\mathcal{R} = (\vec{y} - \Phi\vec{\theta})^T V^{-1} (\vec{y} - \Phi\vec{\theta}) = \sum_{i,j=1}^n (y_i - \theta)V_{ij}^{-1}(y_j - \theta) = \min(\theta)$$

$$\frac{\partial \mathcal{R}}{\partial \theta} = \sum_{i,j=1}^n 2\theta V_{ij}^{-1} - y_i V_{ij}^{-1} - y_j V_{ij}^{-1} = 2\theta \sum_{i,j=1}^n V_{ij}^{-1} - 2 \sum_{i,j=1}^n y_i V_{ij}^{-1}$$

MNK - przypadek liniowy - pomiary skorelowane

$$\left. \frac{\partial \mathcal{R}}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0 \Rightarrow \hat{\theta} = \frac{\sum_{i,j=1}^n y_i V_{ij}^{-1}}{\sum_{k,l=1}^n V_{kl}^{-1}} = \sum_{i=1}^n w_i y_i, \quad w_i = \frac{\sum_{j=1}^n V_{ij}^{-1}}{\sum_{k,l=1}^n V_{kl}^{-1}}$$

$$\langle \hat{\theta} \rangle = \sum_{i=1}^n w_i \langle y_i \rangle = \theta \sum_{i=1}^n w_i = \theta$$

$$\mathcal{V}[\hat{\theta}] = \sum_{i,j=1}^n \frac{\partial \hat{\theta}}{\partial y_i} \frac{\partial \hat{\theta}}{\partial y_j} V_{ij} = \sum_{i,j=1}^n w_i V_{ij} w_j = w^T V w$$

- Przykład: Niech y_1 i y_2 będą dwoma skorelowanymi pomiarami wielkości θ :

$$V = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \Rightarrow V^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix}$$

$$\hat{\theta} = w y_1 + (1-w) y_2, \quad w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$\mathcal{V}[\hat{\theta}] = w^T V w = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

Uwaga 1: Dla $\rho > \sigma_2/\sigma_1$ mamy $w < 0 \Rightarrow \hat{\theta} \notin (y_1, y_2)$ - dla silnie dodatnio skorelowanych pomiarów z dużym prawdopodobieństwem oba pomiary leżą po tej samej stronie wartości prawdziwej.

MNK - przypadek liniowy - pomiary skorelowane

Uwaga 2: Dla $\rho = \sigma_1/\sigma_2$ (lub $\rho = \sigma_2/\sigma_1$) waga drugiego z pomiarów jest równa zero i nie wpływa on na średnią.

- Definiując **błąd skorelowany** (np. niepewność świetlności przy pomiarze dwóch przekrojów czynnych w tym samym eksperymencie) jako $\sigma_C^2 = \rho\sigma_1\sigma_2$ wyniki pomiarów można zapisać:

$$\theta = y_1 \pm \sigma'_1 \pm \sigma_C, \quad \theta = y_2 \pm \sigma'_2 \pm \sigma_C$$

gdzie $\sigma_1'^2 = \sigma_1^2 - \sigma_C^2$ oraz $\sigma_2'^2 = \sigma_2^2 - \sigma_C^2$.

Zakładając, że $\sigma_C < \sigma_1, \sigma_2$ (co oznacza, że wagi $w_{1,2} > 0$) można zapisać:

$$\hat{\theta} = \frac{\frac{y_1}{\sigma_1^2 - \sigma_C^2} + \frac{y_2}{\sigma_2^2 - \sigma_C^2}}{\frac{1}{\sigma_1^2 - \sigma_C^2} + \frac{1}{\sigma_2^2 - \sigma_C^2}} = \frac{\frac{y_1}{\sigma_1'^2} + \frac{y_2}{\sigma_2'^2}}{\frac{1}{\sigma_1'^2} + \frac{1}{\sigma_2'^2}}$$

$$\mathcal{V}[\hat{\theta}] = \frac{1}{\frac{1}{\sigma_1^2 - \sigma_C^2} + \frac{1}{\sigma_2^2 - \sigma_C^2}} + \sigma_C^2 = \frac{1}{\frac{1}{\sigma_1'^2} + \frac{1}{\sigma_2'^2}} + \sigma_C^2$$

- Przykład: Wykorzystując dwa razy ten sam pomiar i zaniehbując korelację zmniejszamy błąd o $\sqrt{2}$. Jeśli poprawnie przyjmiemy, że wówczas $\rho = 1$ to oczywiście taka operacja nie zmniejsza błędu.