

Uczenie maszynowe w bioinformatyce

Wykład 7: techniki redukcji wymiarowości (na przykładzie analizy składowych głównych)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Etapy przetwarzania i analizy danych

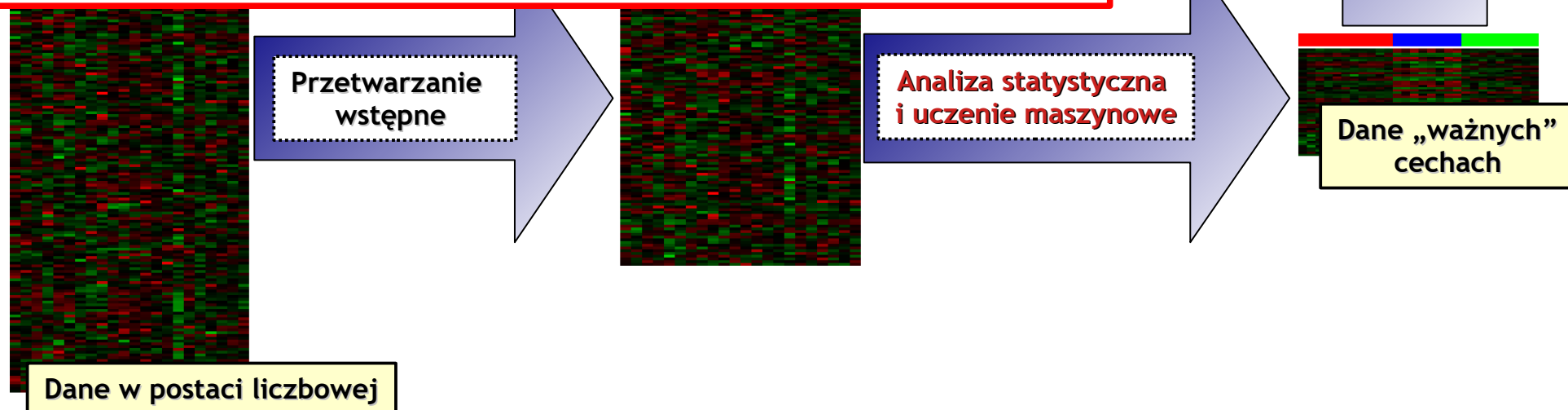
Wybór **metod statystycznych i uczenia się maszyn** używanych na etapie analizy danych jest ściśle uzależniony od celu badań. Stosowane tutaj techniki można w ogólności podzielić na **działające z nadzorem** lub **nienadzorowane**.

Metody działające bez nadzoru stosuje się do wykrywania zależności pomiędzy próbkami i/lub cechami jedynie w oparciu o dane, bez używania dodatkowych informacji (w tym też nie uwzględniając przynależności badanych obiektów do zadanych z góry klas). Do metod nienadzorowanych zaliczają się m. in.:

- ✓ klasteryzacja;
- ➔ techniki redukcji wymiarowości.

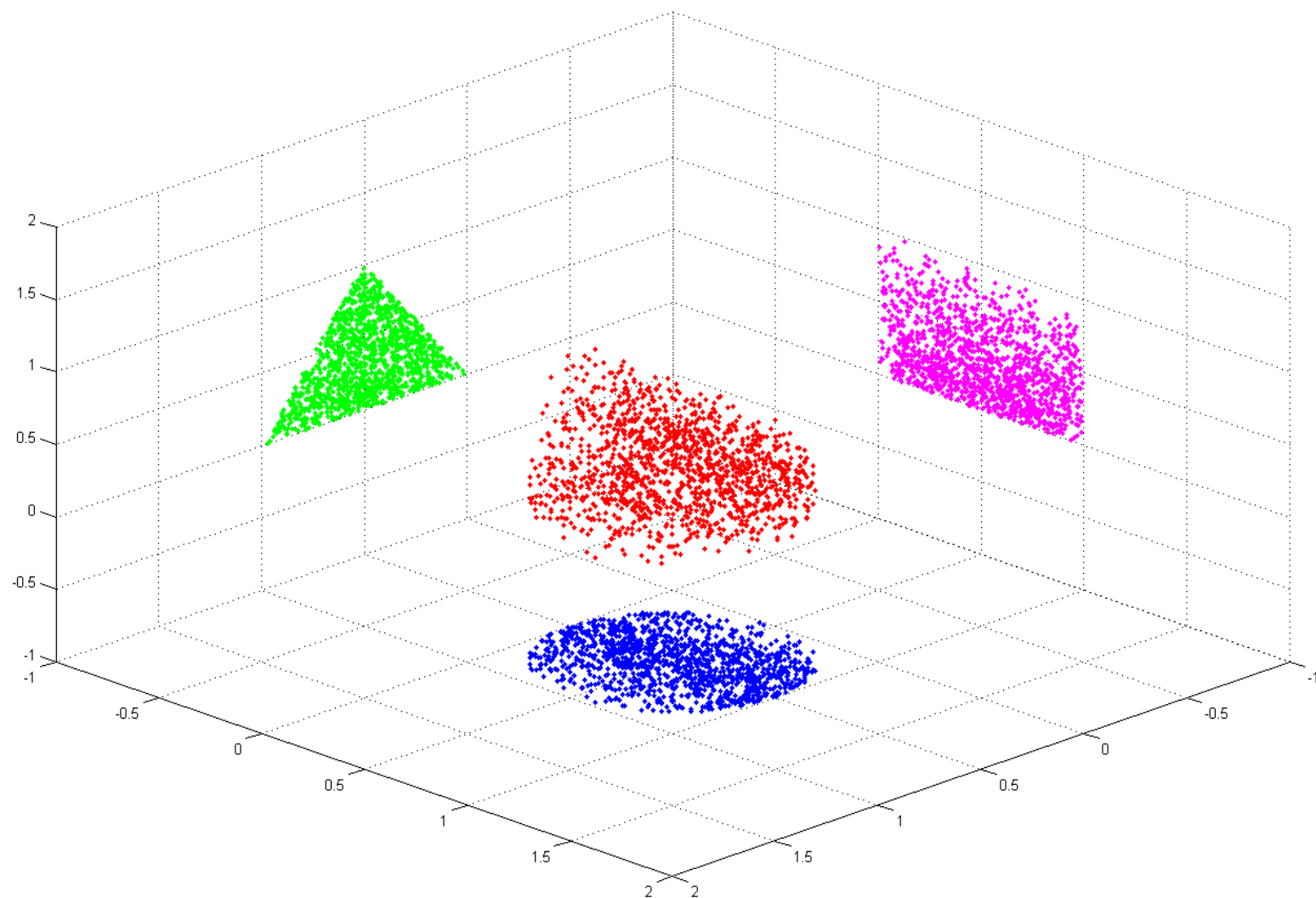
Metody nadzorowane korzystają z danych i dodatkowej wiedzy, dotyczącej np. oczekiwanych wartości na wyjściu lub sposobu przypisania próbek do znanych klas. Przykładami zagadnień, w których używa się takich metod są:

- klasyfikacja;
- selekcja cech różnicujących grupy próbek.



Redukcja wymiarowości

Redukcją wymiarowości określa się wyznaczanie zbioru cech, mniej licznego od zbioru cech pierwotnych, na podstawie których można z jak najmniejszym błędem odtworzyć wartości pierwotnych cech. Jest to równoznaczne ze znalezieniem nowej przestrzeni o niższej wymiarowości, wybranej w taki sposób, aby przerzutowanie do niej danych wiązało się z możliwie małą utratą informacji.

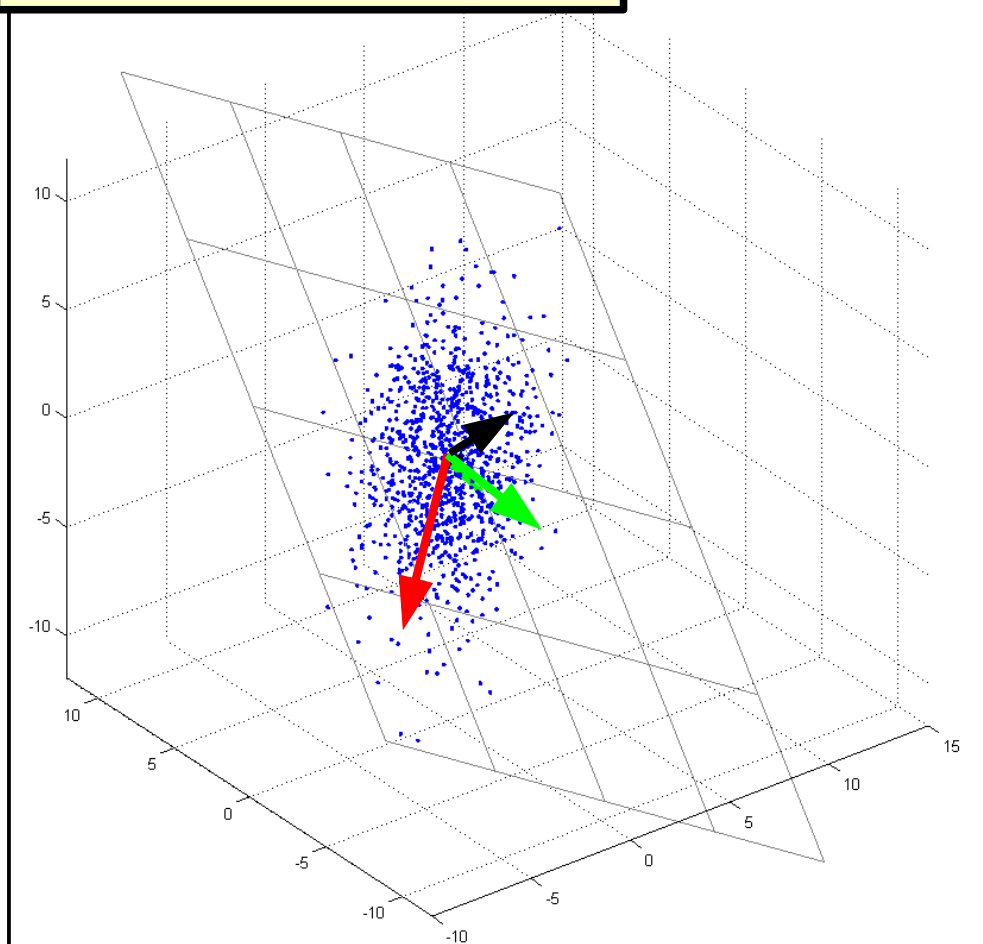


Redukcja wymiarowości: analiza składowych głównych

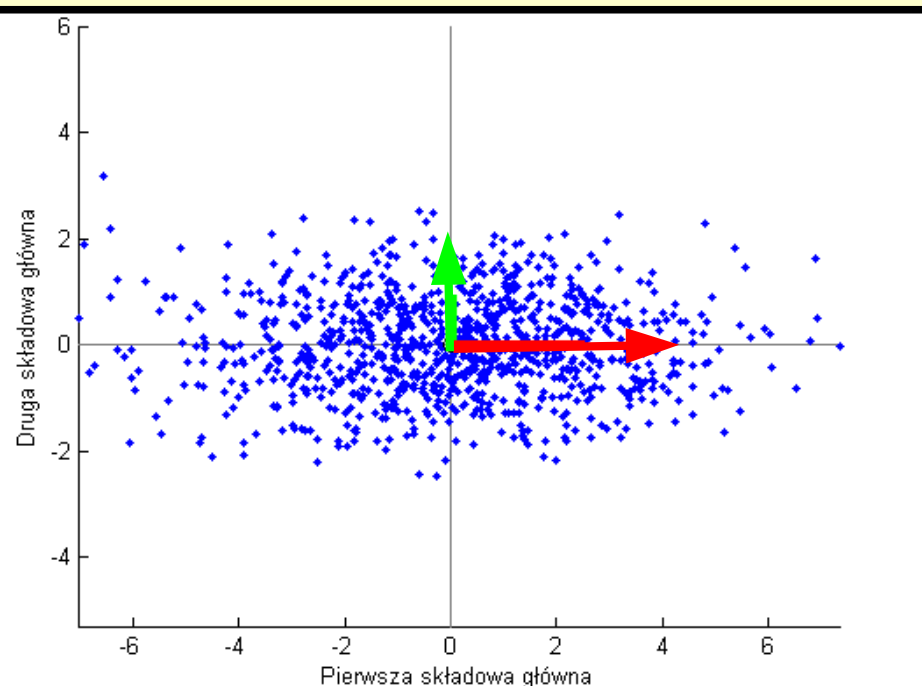
Jedną z najpopularniejszych technik redukcji wymiarowości jest **analiza składowych głównych (PCA – Principal Components Analysis)**.

PCA polega na przerzutowaniu danych do przestrzeni (o takiej samej lub mniejszej wymiarowości), w której osie układu współrzędnych, zwane **składowymi głównymi**, skierowane są zgodnie z kierunkami największej wariancji danych w oryginalnej przestrzeni cech.

Dane w przestrzeni oryginalnych cech



Dane w przestrzeni dwóch pierwszych składowych głównych



	MATLAB	R
PCA:	pca/princomp	prcomp/princomp

Redukcja wymiarowości: analiza składowych głównych

Wartości każdej z R ($R \leq P$) cech w przestrzeni składowych głównych stanowią kombinacje liniowe wartości wszystkich P oryginalnych cech. Współczynnikami kombinacji są współrzędne ortonormalnych wektorów (czyli wektorów wzajemnie prostopadłych o długości 1), które wyznaczają osie nowego układu współrzędnych:

$$\left. \begin{aligned} y_1 &= t_{11}x_1 + t_{12}x_2 + \dots + t_{1P}x_P = t_1^T X \\ y_2 &= t_{21}x_1 + t_{22}x_2 + \dots + t_{2P}x_P = t_2^T X \\ &\dots \\ y_R &= t_{R1}x_1 + t_{R2}x_2 + \dots + t_{RP}x_P = t_R^T X \end{aligned} \right\} Y = T^T X$$

x_j – wektory wartości j -tej cechy w pierwotnej przestrzeni (j -ty wiersz macierzy X)

t_i – i -ta składowa główna (wiersz i -tej osi nowego układu współrzędnych)

y_k – wektory wartości k -tej cechy w przestrzeni składowych głównych (k -ty wiersz macierzy Y , będącej projekcją danych w przestrzeni składowych głównych)

Pierwsza składowa główna t_1 wyznaczana jest tak, aby uzyskać kombinację liniową o maksymalnej wariancji z próby spośród wszystkich kombinacji liniowych wartości cech pierwotnych (przy zachowaniu długości wektora t_1 równej 1).

Każda kolejna składowa główna t_i ($i=2, \dots, R$) wyznaczana jest jako wektor o długości 1, ortogonalny do pozostałych składowych i maksymalizujący wariancję z próby odpowiadającej mu kombinacji liniowej.

Redukcja wymiarowości: analiza składowych głównych

Dla macierzy X o wymiarze $P \times N$, rzędzie R i zerowym wektorze wartości średnich, **wektorami t rozpinającymi przestrzeń składowych głównych są wektory własne macierzy kowariancji z próby S_X , odpowiadające uporządkowanym nierosnąco, niezerowym wartościom własnym: $\lambda_1 \geq \lambda_2 \geq \dots \lambda_R \geq 0$.**

Macierz kowariancji z próby:
$$S = \hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{N-1} (X - \bar{x})(X - \bar{x})^T$$

Gdy wektor wartości średnich macierzy X jest równy 0:
$$S = \frac{1}{N-1} \sum_{i=1}^N x_i x_i^T = \frac{1}{N-1} X X^T$$

Wartością własną macierzy S nazywamy każdy skalar λ spełniający równanie:

$$|S - \lambda I| = 0$$

gdzie I jest diagonalną macierzą jednostkową o wymiarach takich samych jak S .

Wektorem własnym macierzy S nazywamy każdy wektor spełniający równanie:

$$S t = \lambda t$$

Redukcja wymiarowości: związek pomiędzy PCA i SVD

Dla danych z mikromacierzy liczenie składowych głównych na podstawie macierzy kowariancji jest utrudnione przez wymiarowość tej ostatniej ($P \times P$, a P może wynosić kilkadziesiąt tysięcy). Dlatego też składowe główne zwykle wyznacza się poprzez **rozkład macierzy X na wartości szczególne (SVD – Singular Value Decomposition)**.

Macierz X o wymiarze $P \times N$ i rzędzie R można zapisać w postaci rozkładu:

$$X = U \Sigma V^T$$

Σ – diagonalna prostokątna macierz o wymiarach $P \times N$, na przekątnej której znajdują się rzeczywiste, nieujemne elementy będące uporządkowanymi niemalejąco wartościami szczególnymi macierzy X : $\sigma_1 \geq \sigma_2 \geq \dots \sigma_R \geq \sigma_{R+1} = \dots = \sigma_{\min(N,P)} = 0$

U – ortonormalna macierz o wymiarach $P \times P$, której pierwsze R kolumn stanowi lewostronne wektory szczególne macierzy X

V – ortonormalna macierz o wymiarach $N \times N$, której pierwsze R kolumn stanowi prawostronne wektory szczególne macierzy X

Lewostronne wektory własne są wektorami własnymi macierzy $X \cdot X^T$, zaś wartości szczególne są pierwiastkami wartości własnych tej macierzy. Gdy wektor wartości średnich macierzy X jest równy 0 (co jest wymagane dla PCA), to macierz $X \cdot X^T$ jest równa macierzy kowariancji S_X pomnożonej przez $(N - 1)$, z czego wynika że:

$$T = U$$

$$\lambda_i = \sigma_i^2 / (N - 1)$$

Rozkład SVD:

MATLAB

svd

R

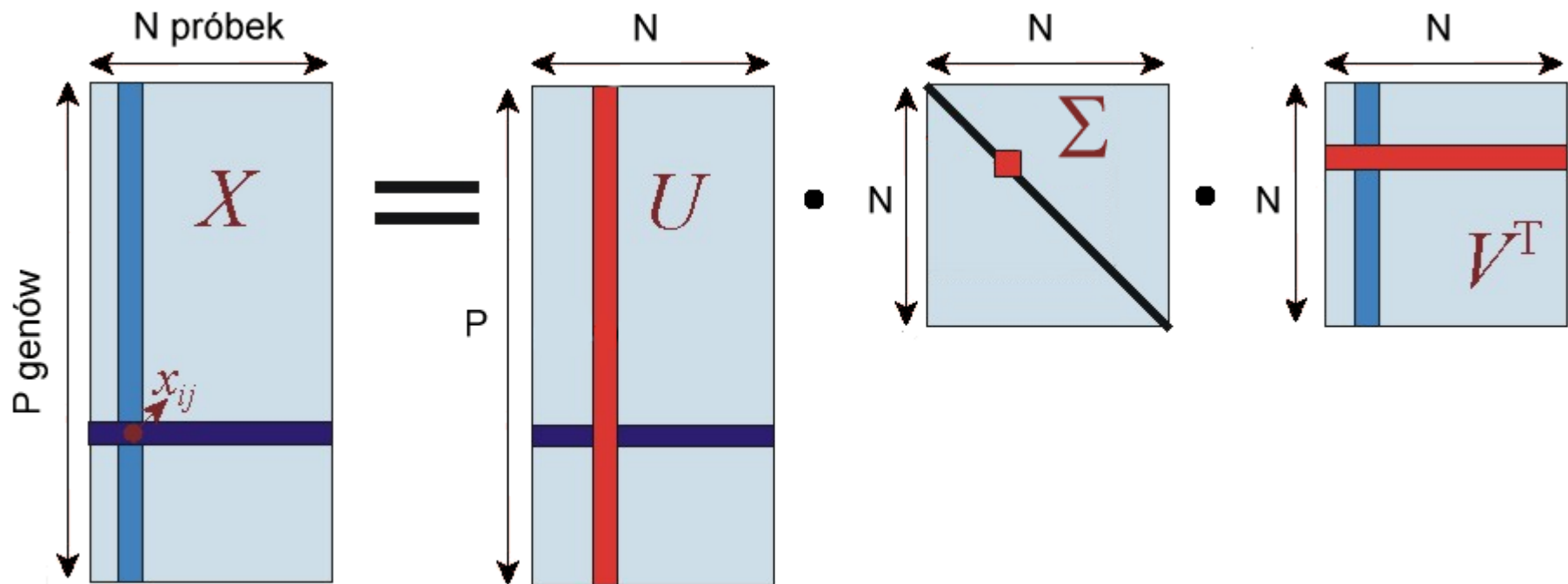
svd

Redukcja wymiarowości: związek pomiędzy PCA i SVD

W szczególnym przypadku, gdy dla macierzy X liczba wierszy P jest większa liczby kolumn N , to macierz Σ z rozkładu SVD przyjmuje postać:

$$\begin{bmatrix} \Sigma_{N \times N} \\ \mathbf{0}_{(P-N) \times N} \end{bmatrix}$$

Prowadzi to do postaci rozkładu macierzy X o mniejszym wymaganiach dotyczących pamięci operacyjnej:



Redukcja wymiarowości: właściwości analizy składowych głównych

- Jako transformata ortonormalna **nie zmienia długości wektorów i odległości Euklidesa pomiędzy nimi**:

$$d(x_1, x_2) = d(T^T x_1, T^T x_2) = d(y_1, y_2)$$

$$\|x\| = \|T^T x\| = \|y\|$$

- **Cechy w przestrzeni składowych głównych są nieskorelowane** (kowariancja każdej pary cech wynosi 0). W efekcie macierz kowariancji z próby cech S_Y jest diagonalna:

$$S_Y = T^T S_X T = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_R \end{bmatrix}$$

- **Wariancja danych w kierunku i -tej składowej głównej jest równa i -tej wartości własnej λ_i** . Całkowita wariancja z próby cech w przestrzeni składowych głównych równa jest sumie wszystkich wartości własnych i jednocześnie równa sumarycznej wariancji z próby cech w pierwotnej przestrzeni:

$$\sum_{i=1}^R \lambda_i = \text{tr}(S_Y) = \sum_{j=1}^P s_j^2 = \text{tr}(S_X)$$

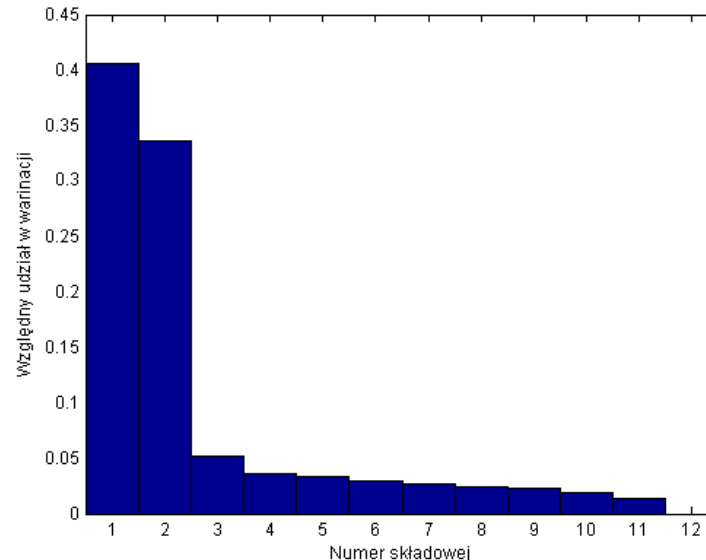
Redukcja wymiarowości: właściwości analizy składowych głównych

- Uporządkowanie wartości własnych powoduje, że wariancji w kierunku pierwszej składowej jest nie mniejsza od wariancji z próby w kierunku jakiegokolwiek innej składowej (jednocześnie jest ona nie mniejsza od wariancji z próby jakiegokolwiek innej kombinacji liniowej pierwotnych cech). Oznacza to, że **kolejne składowe są uporządkowane zgodnie ze swoim udziałem w wariancji zbioru danych**.

Względny udział i -tej składowej głównej w wariancji całego zbioru danych można wyrazić jako:

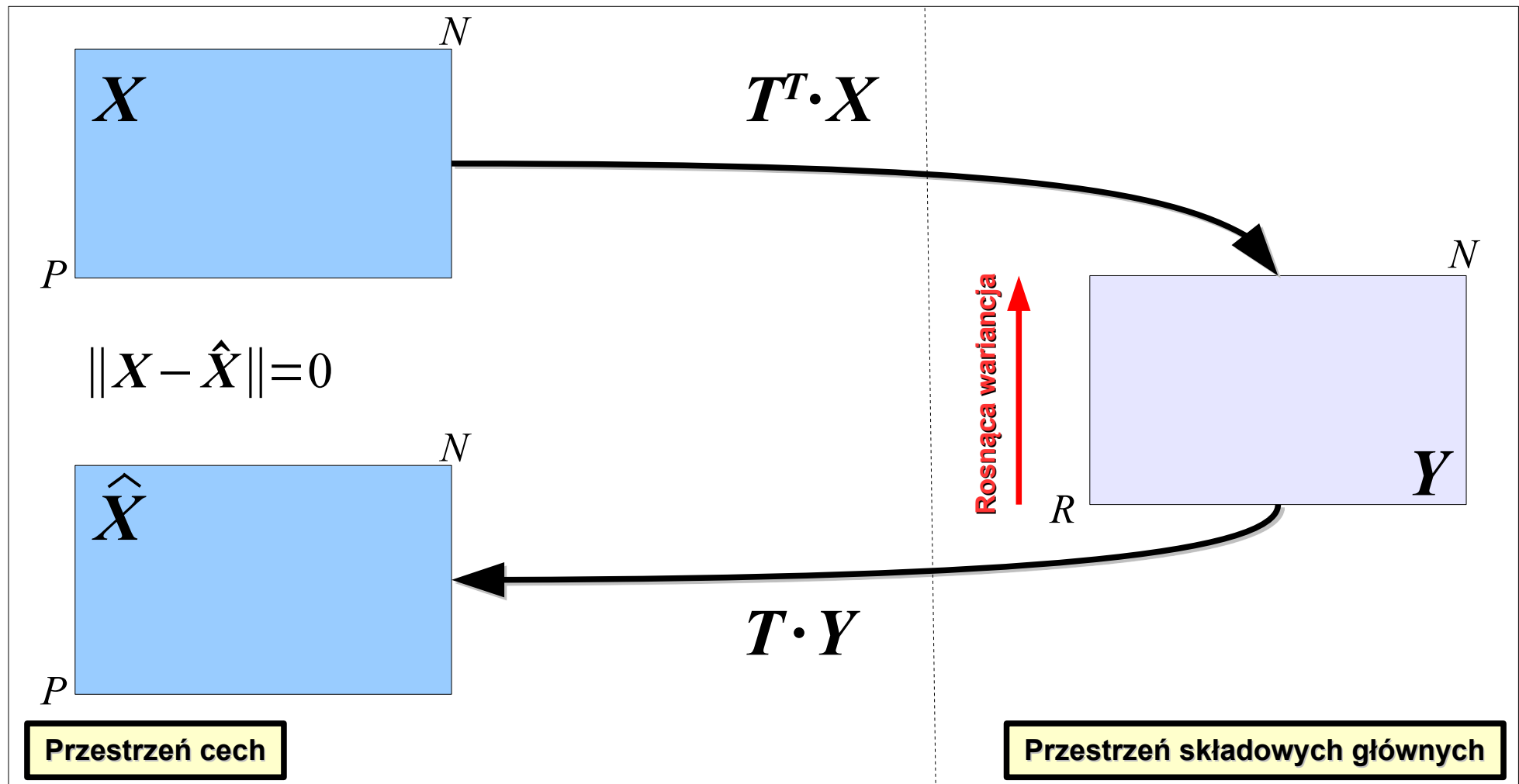
$$p_i = \frac{\lambda_i}{\sum_{j=1}^R \lambda_j}$$

Zazwyczaj kilka pierwszych składowych „przechwytuje” prawie całą wariancję zbioru danych, podczas gdy wkład pozostałych jest zdecydowanie mniejszy.



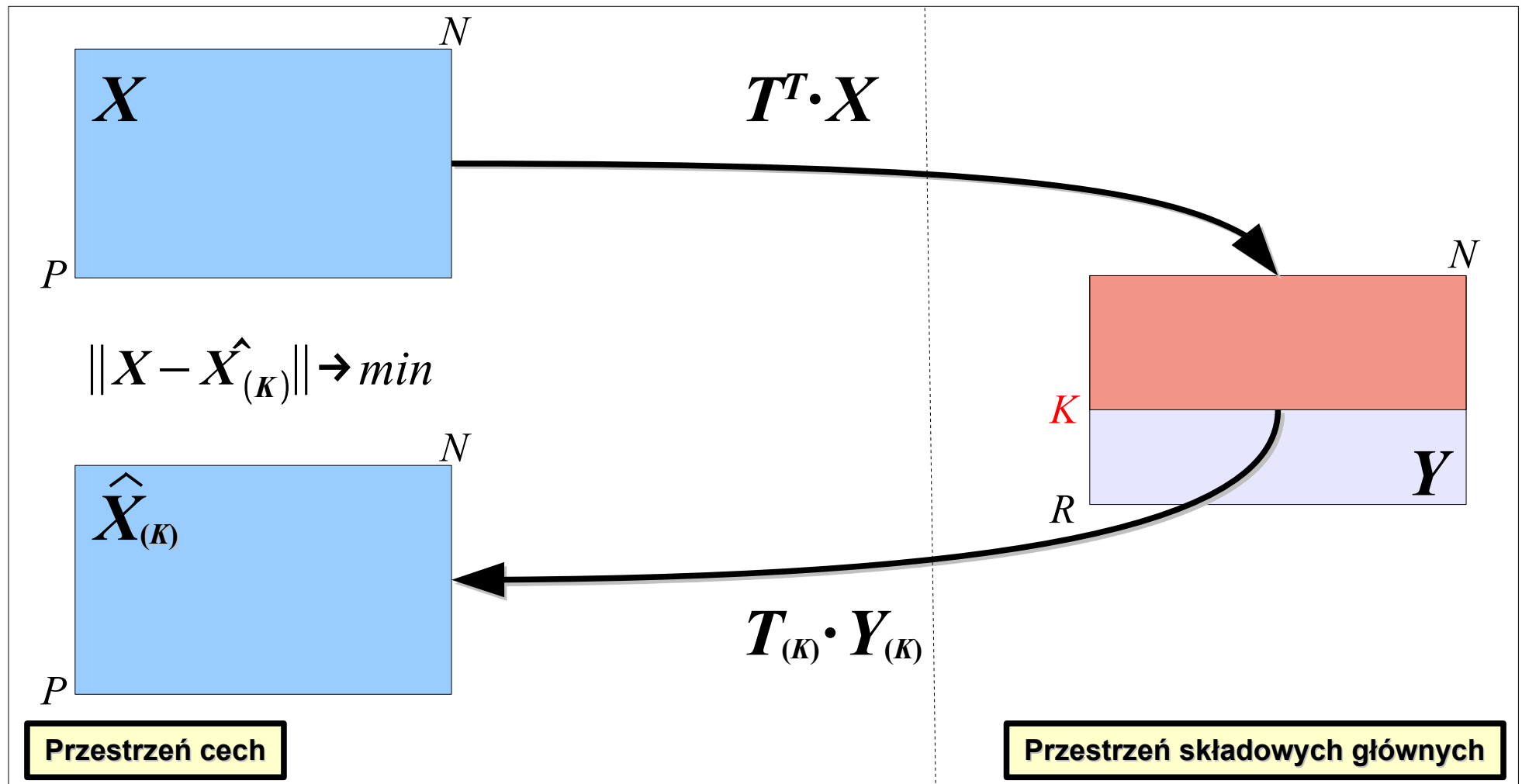
Redukcja wymiarowości: właściwości analizy składowych głównych

Właściwości te można podsumować stwierdzeniem, że **PCA pozwala opisać dane o wymiarowości P za pomocą R ($R \leq P$) nieskorelowanych cech uporządkowanych pod względem ich udziału w całkowitej wariancji.**



Redukcja wymiarowości: właściwości analizy składowych głównych

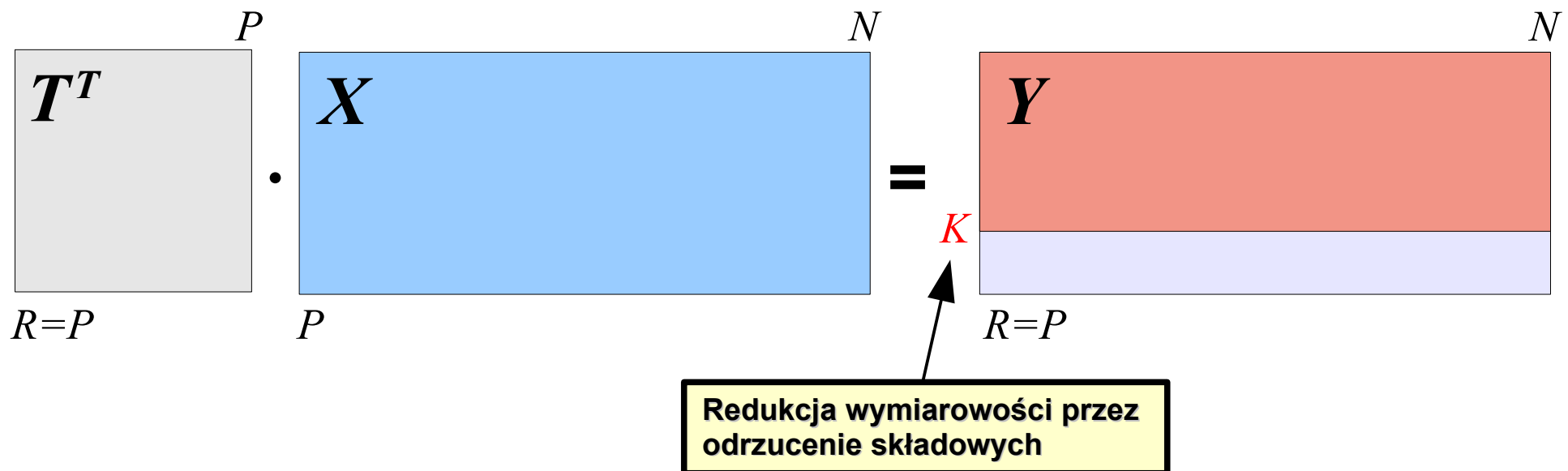
W praktyce najczęściej pierwszych K ($K \leq R$) składowych „przechwytuje” niemal całą wariancję oryginalnych cech. Oznacza to, że można zastąpić P pierwotnych cech zbiorem K cech w przestrzeni składowych głównych, minimalizując przy tym utratę informacji wynikającą z redukcji wymiaru.



Redukcja wymiarowości: właściwości analizy składowych głównych

Składowych głównych jest tyle, ile istnieje niezerowych wartości własnych macierzy kowariancji S_x , czyli tyle ile wynosi rząd R tej macierzy. Ten ostatni zaś równy jest co najwyżej $\min(P, N-1)$, gdzie P jest liczbą wierszy macierzy danych (liczbą cech) a N to liczba kolumn (liczba próbek).

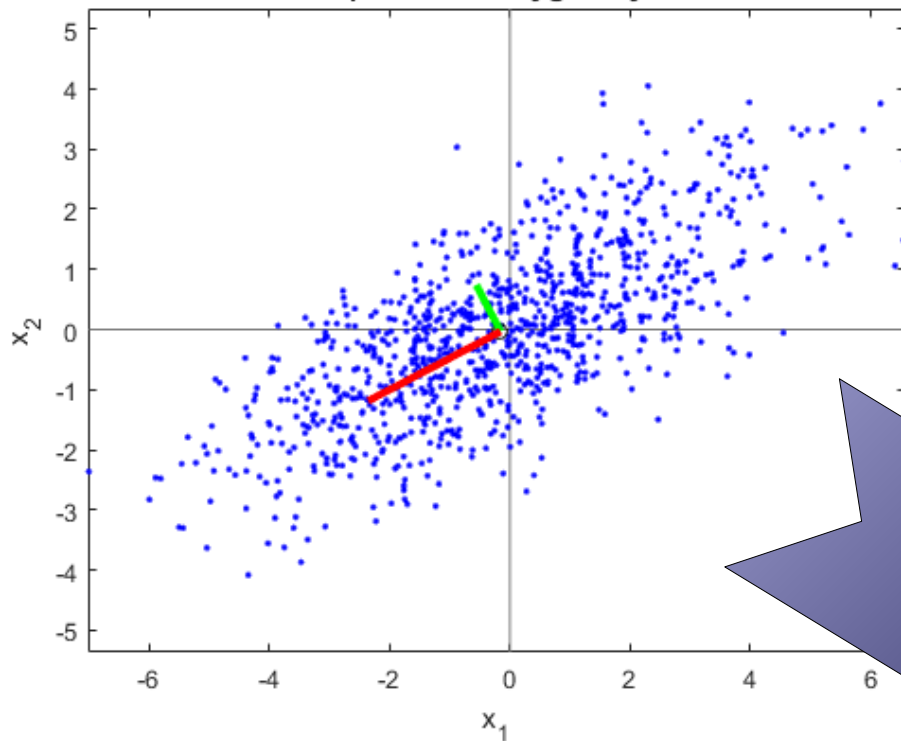
Dla „zwykłych” danych, dla których liczba próbek jest większa od liczby cech ($P < N$), rząd macierzy kowariancji będzie zwykle równy liczbie cech ($R = P$). Czyli samo przeniesienie danych do przestrzeni składowych głównych nie powoduje redukcji ich wymiaru. **Zmniejszenie wymiarowości jest możliwe dopiero po wybraniu pewnej liczby K początkowych składowych, niosących największą wariancję.**



Redukcja wymiarowości: właściwości analizy składowych głównych

Przykład dla danych dwuwymiarowych ($P = 2, N = 1000$):

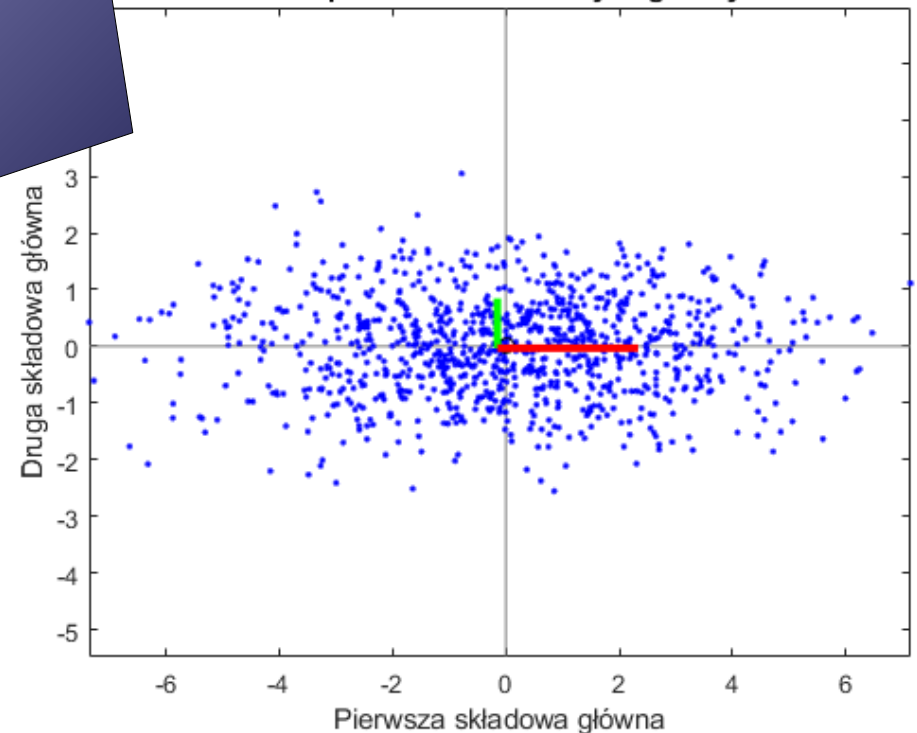
Dane w przestrzeni oryginalnych cech



Przerzutowanie do przestrzeni składowych głównych powoduje zmianę układu współrzędnych, na taki, którego osie skierowanie są zgodnie z kierunkami największego rozrzutu danych.

Redukcja wymiarowości jest możliwa po odrzuceniu drugiej składowej głównej i wykonanie rzutu danych na oś wyznaczoną przez pierwszą składową.

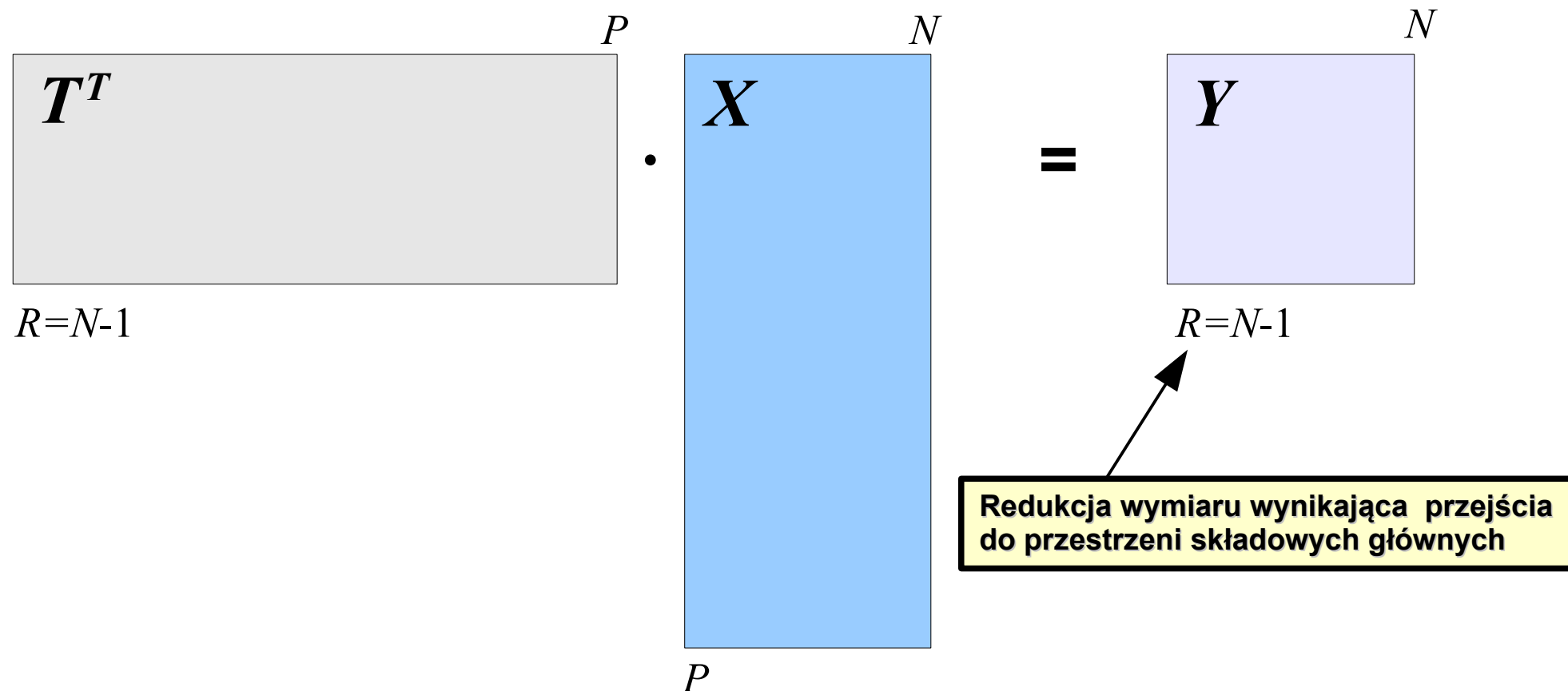
Dane w przestrzeni składowych głównych



Redukcja wymiarowości: właściwości analizy składowych głównych

Składowych głównych jest tyle, ile istnieje niezerowych wartości własnych macierzy kowariancji S_x , czyli tyle ile wynosi rząd R tej macierzy. Ten ostatni równy jest co najwyżej $\min(P, N-1)$, gdzie P jest liczbą wierszy macierzy danych (liczbą cech) a N jest liczbą kolumn (liczbą próbek).

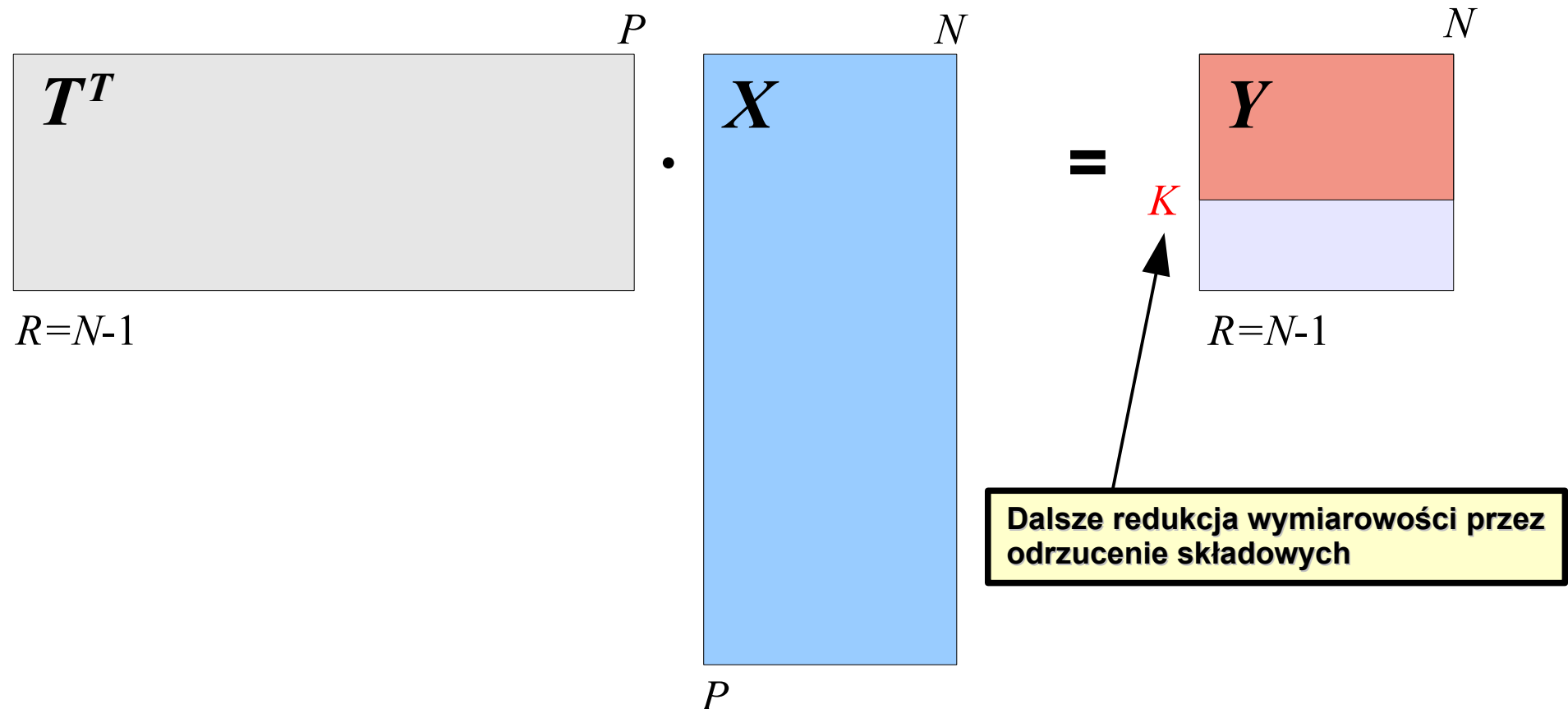
Dla danych z mikromacierzy $P \gg N$, a rząd macierzy kowariancji wynosi co najwyżej $N-1$. W efekcie, już samo przejście do przestrzeni składowych głównych powoduje znaczącą redukcję wymiaru (z P do $N-1$) przy zachowaniu całości wariancji danych.



Redukcja wymiarowości: właściwości analizy składowych głównych

Składowych głównych jest tyle, ile istnieje niezerowych wartości własnych macierzy kowariancji S_x , czyli tyle ile wynosi rząd R tej macierzy. Ten ostatni równy jest co najwyżej $\min(P, N-1)$, gdzie P jest liczbą wierszy macierzy danych (liczbą cech) a N jest liczbą kolumn (liczbą próbek).

Dla danych z mikromacierzy $P \gg N$, a rząd macierzy kowariancji wynosi co najwyżej $N-1$. W efekcie, już samo przejście do przestrzeni składowych głównych powoduje znaczącą redukcję wymiaru (z P do $N-1$) przy zachowaniu całości wariancji danych.



Redukcja wymiarowości: zastosowania PCA

Zastosowania analizy składowych głównych:

- **zmniejszenie liczby cech przed właściwą analizą danych** przy użyciu innej metody, przykładowo przed klasteryzacją lub klasyfikacją;
- jedno-, dwu- lub trójwymiarowa **wizualizacja danych** wielowymiarowych;
- **redukcja poziomu szumu** w danych;
- **imputacja brakujących wartości**;
- **selekcja cech**, np. selekcja grup genów o wspólnych wzorcach ekspresji;
- wiele innych ...

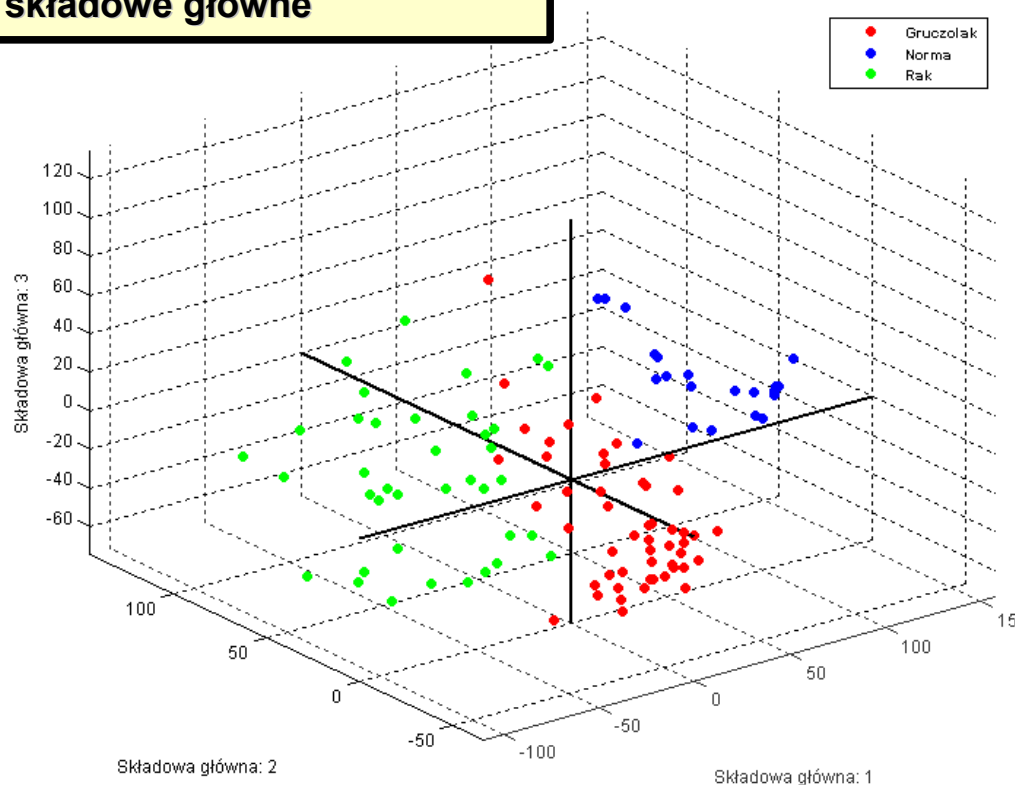
Redukcja wymiarowości: zastosowania PCA (wizualizacja)

Płaszczyzna wyznaczona przez dwie pierwsze składowe główne może posłużyć do dwuwymiarowej **wizualizacji zależności pomiędzy próbkami w zbiorze danych**. Rzut danych na płaszczyznę uzyskuje się uwzględniając jedynie dwie pierwsze kolumny macierzy transformacji T (dwa pierwsze wektory własne):

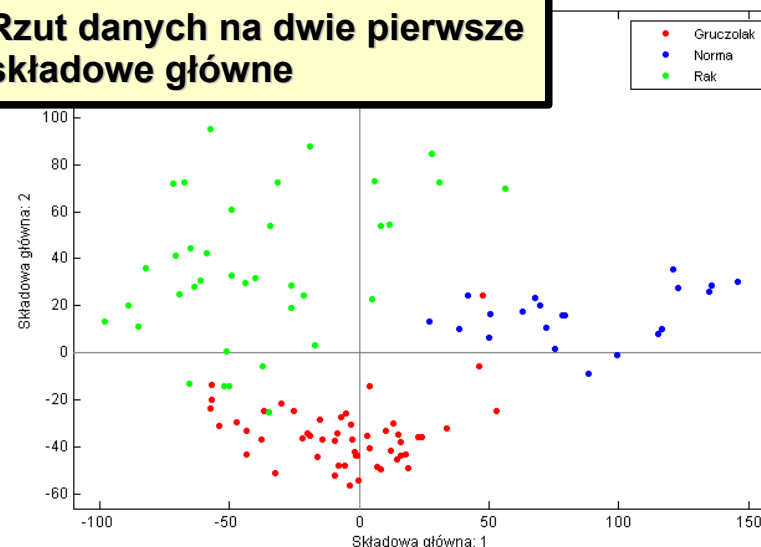
$$Y_2 = T_2^T X = [t_1 t_2]^T X$$

W analogiczny sposób uzyskuje się rzuty jedno- i trójwymiarowe.

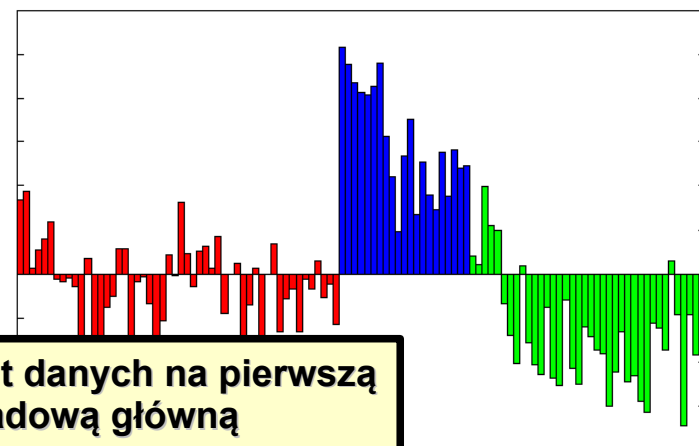
Rzut danych na trzy pierwsze składowe główne



Rzut danych na dwie pierwsze składowe główne



Rzut danych na pierwszą składową główną



Redukcja wymiarowości: zastosowania PCA (redukcja poziomu szumu)

Redukcja szumu przy użyciu PCA opiera się na założeniu, że całość zawartych w danych informacji związane jest z K pierwszymi składowymi głównymi ($K < R$), podczas gdy ostatnie składowe, mające najmniejszy udział w całkowitej wariancji, niosą głównie szum i mogą być pominięte.

Procedura redukcji szumu:

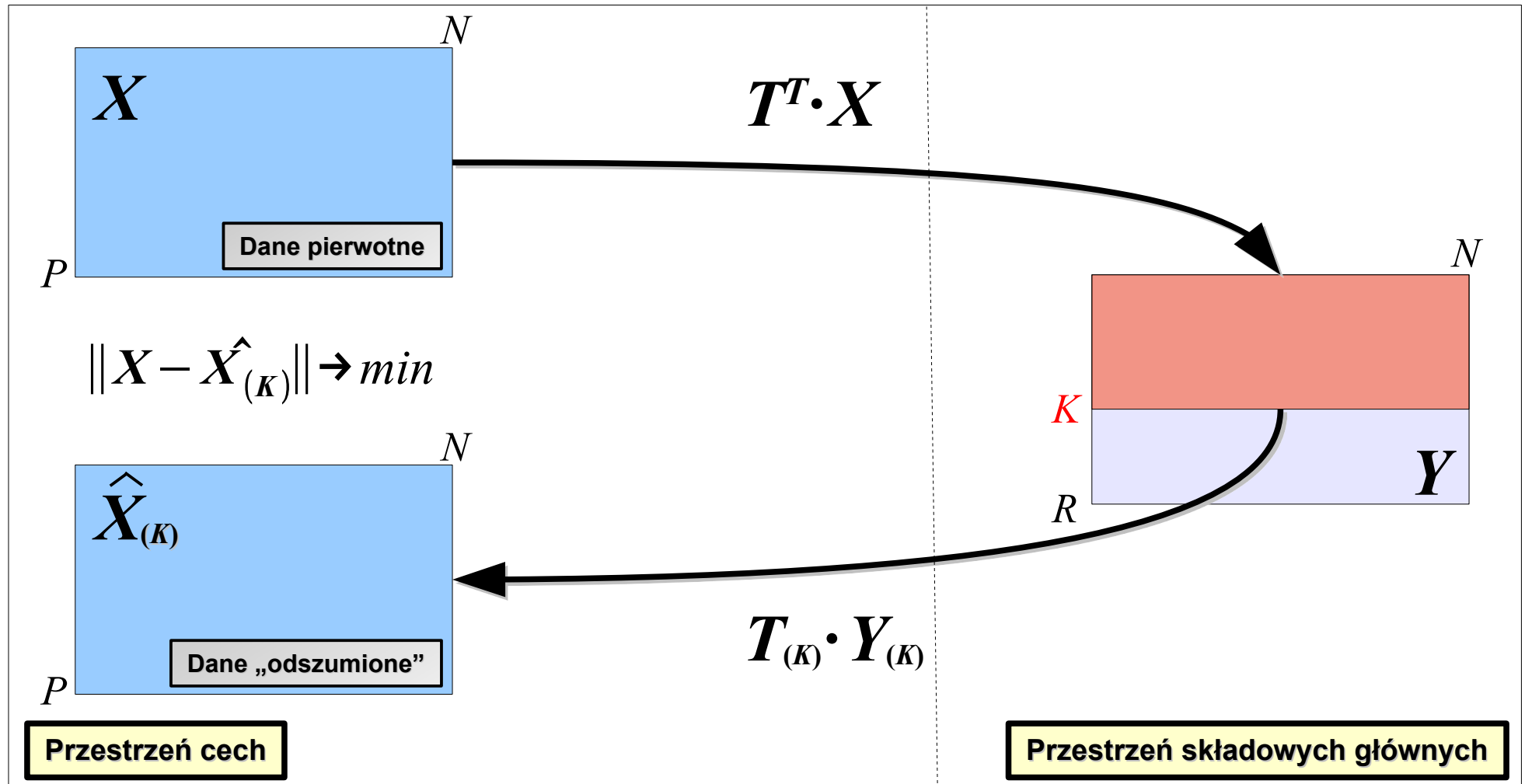
- Wyznaczenie reprezentacji danych w przestrzeni składowych głównych:

$$Y = T^T X$$

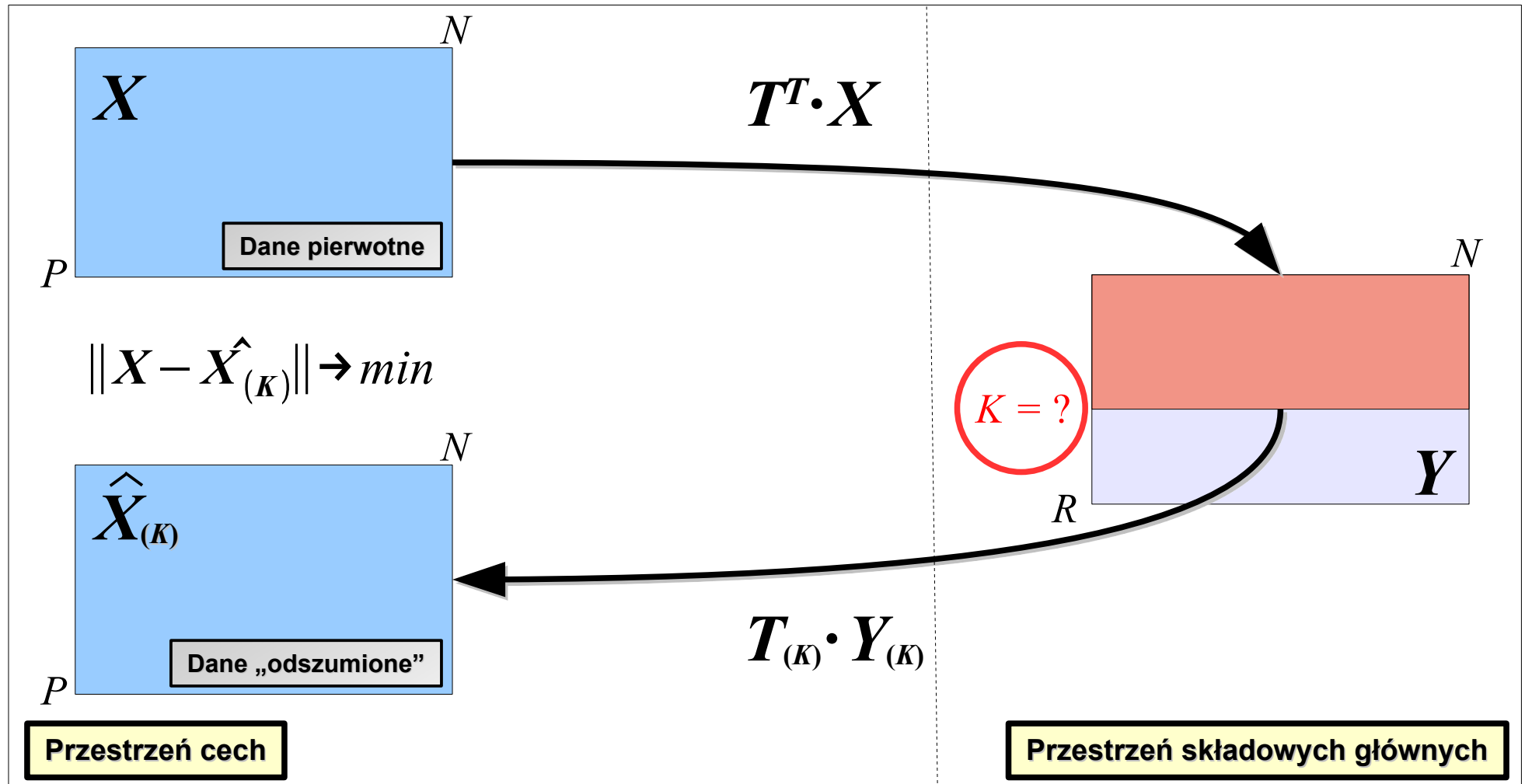
- Wybór K składowych o największym udziale w całkowitej wariancji danych
- Przeniesienie danych z powrotem do oryginalnej przestrzeni odbywa się przy użyciu macierzy T_K , złożonej z wektorów własnych odpowiadających K największym wartościom własnym:

$$\hat{X} = T_K Y = [t_1 t_2 \dots t_K] Y$$

Redukcja wymiarowości: zastosowania PCA (redukcja poziomego szumu)



Redukcja wymiarowości: zastosowania PCA (redukcja poziomego szumu)



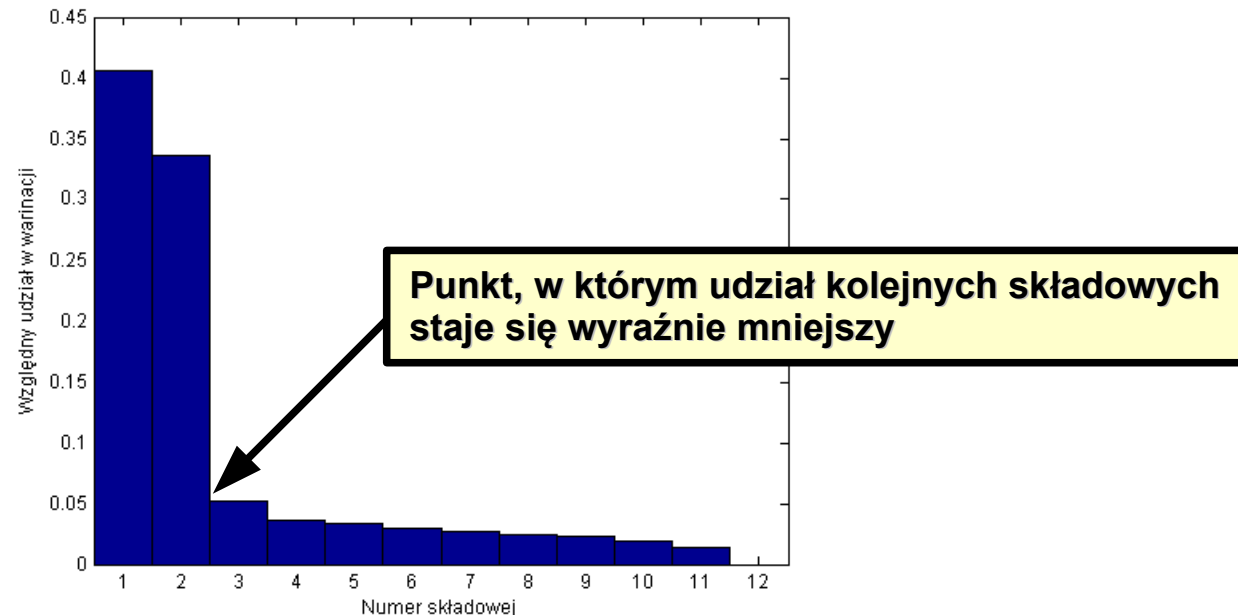
Redukcja wymiarowości: zastosowania PCA (redukcja poziomu szumu)

Przykłady metod wyboru K składowych istotnych dla opisu danych:

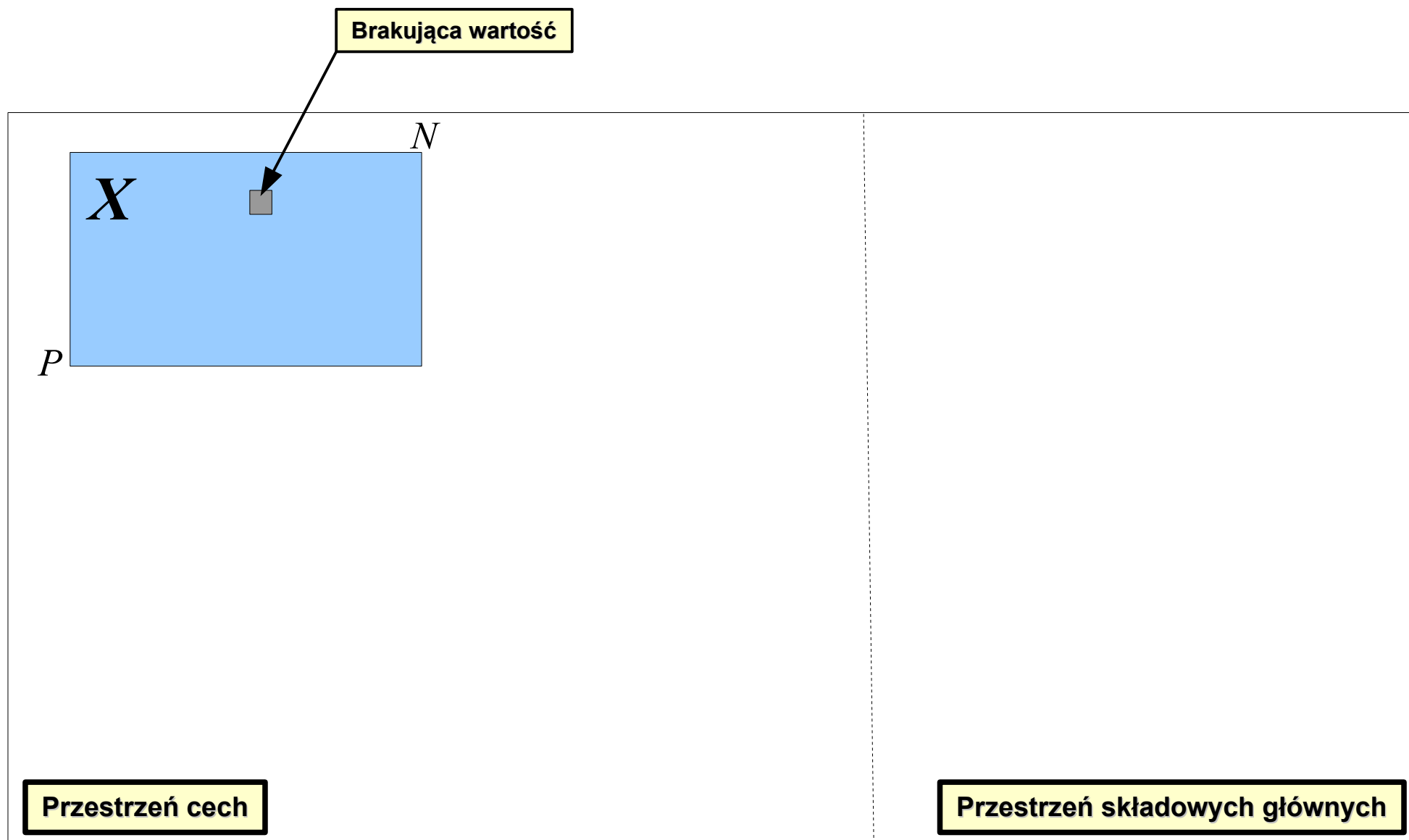
- Wybór K składowych wyjaśniających co najmniej $E\%$ (np. 90%) całkowitej wariancji:

$$\frac{\sum_{i=1}^K \lambda_i}{\sum_{j=1}^R \lambda_j} 100\% \geq E\%$$

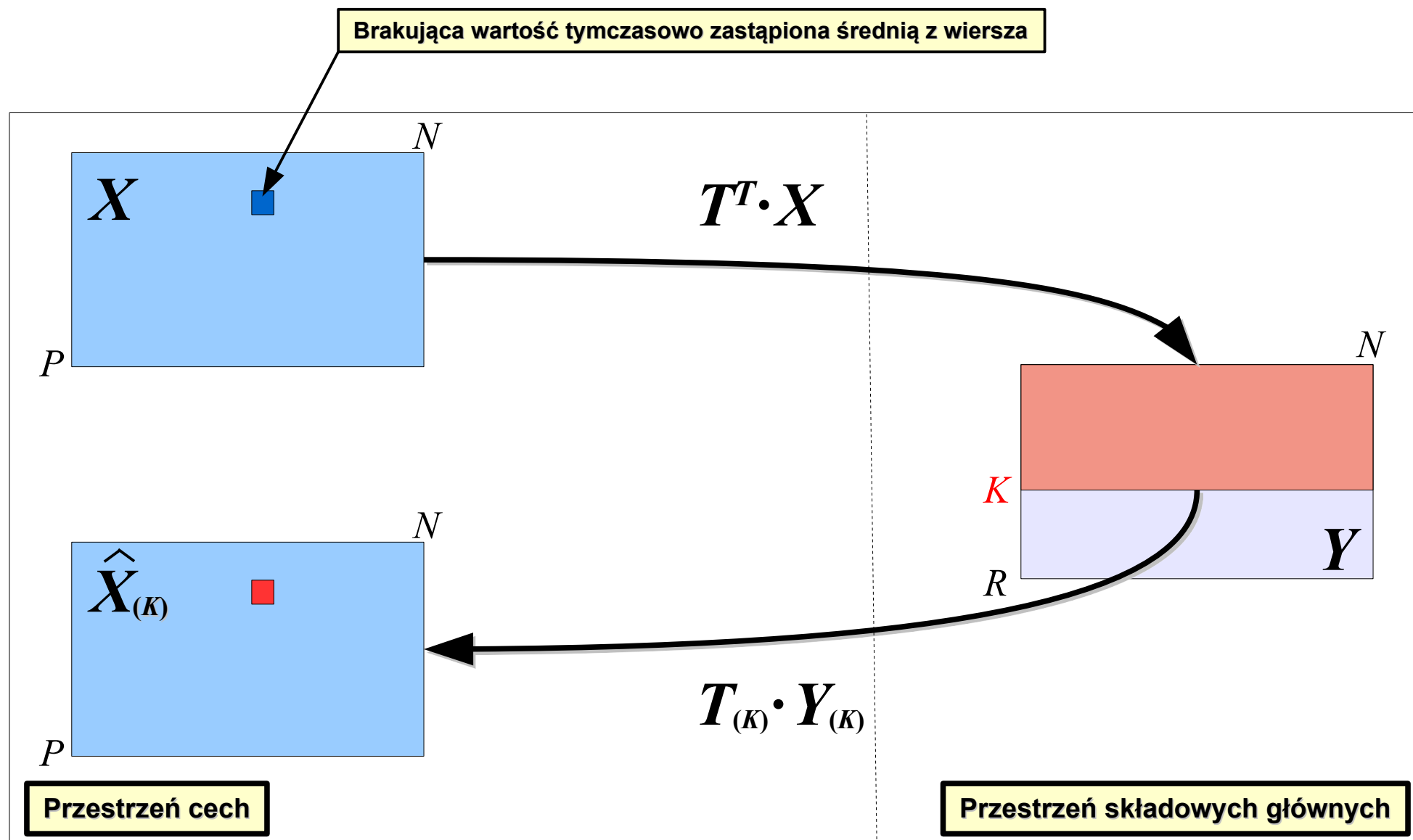
- Wybór na podstawie wykresu względnego udziału składowych w sumarycznej wariancji zbioru danych (tzw. wykresu osypiskowego):



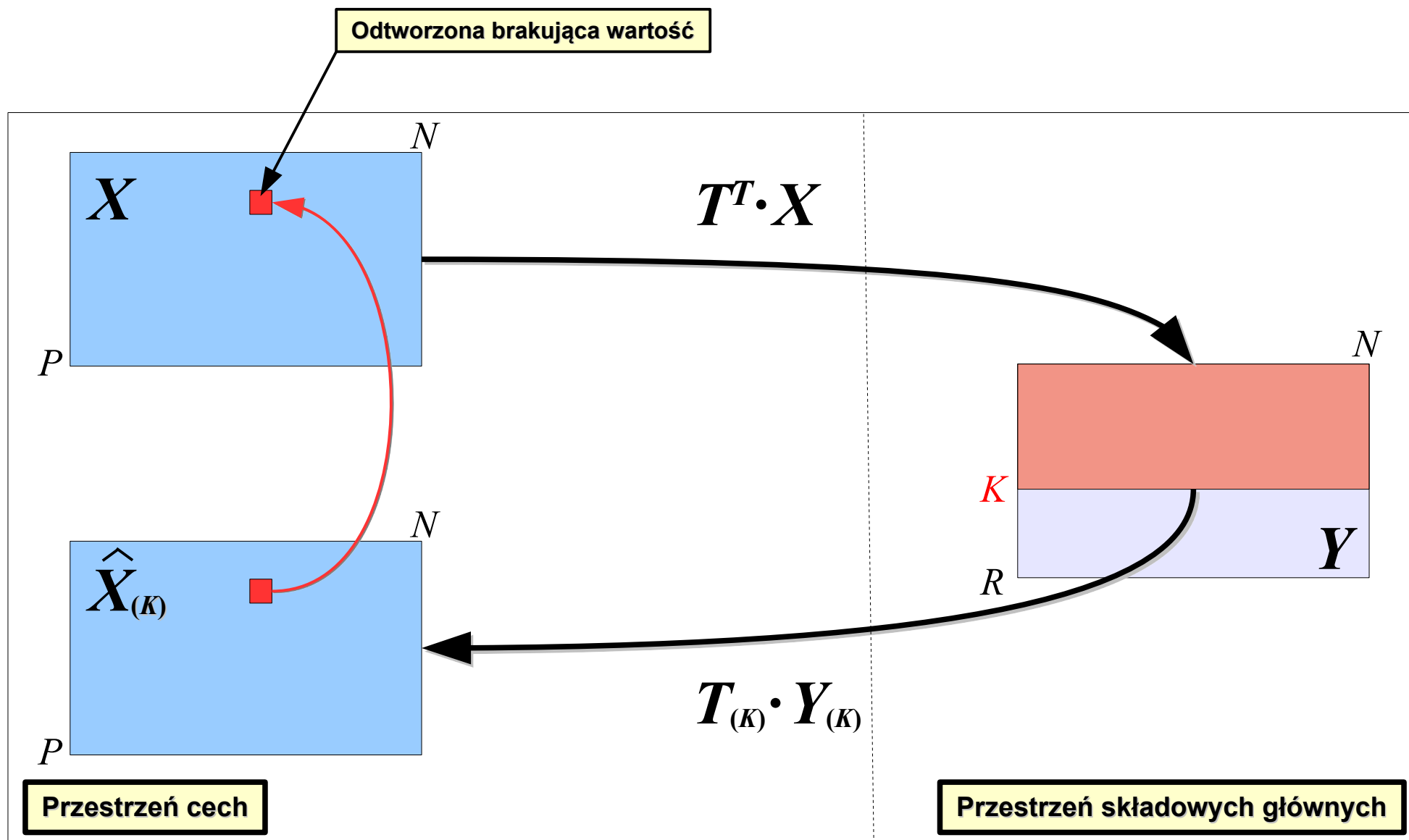
Redukcja wymiarowości: zastosowania PCA (imputacja brakujących wart.)



Redukcja wymiarowości: zastosowania PCA (imputacja brakujących wart.)

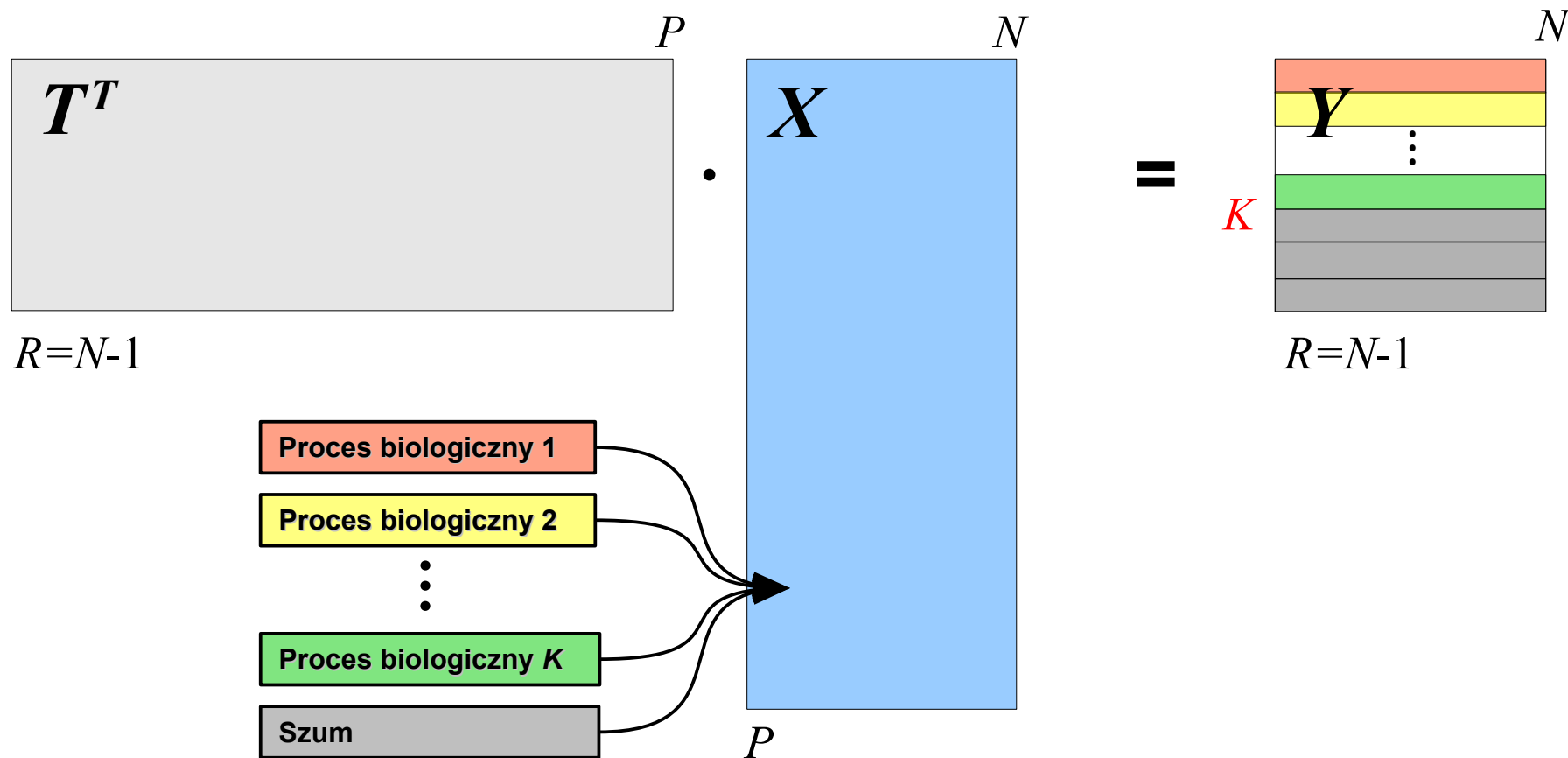


Redukcja wymiarowości: zastosowania PCA (imputacja brakujących wart.)



Redukcja wymiarowości: zastosowania PCA (selekcja cech)

Brak korelacji pomiędzy cechami w przestrzeni składowych głównych powoduje, że możemy próbować je interpretować jako efekt wpływu na ekspresję genów zbioru niezależnych procesów biologicznych. Pozwala to na **selekcję genów o wzorcach najsilniej powiązanych z procesami reprezentowanymi przez kolejne składowe**.



Redukcja wymiarowości: zastosowania PCA (selekcja cech)

Dwa podejścia do selekcji genów związanych z i -tą składową główną:

- Wartości i -tej cechy w przestrzeni składowych głównych są kombinacjami liniowymi wartości ekspresji wszystkich genów. Wybieramy geny mające największy udział w powstaniu i -tej cechy w przestrzeni składowych głównych, czyli geny, dla których wartości bezwzględne współczynników kombinacji liniowej tworzącej i -tą cechę są większe od zadanego progu.

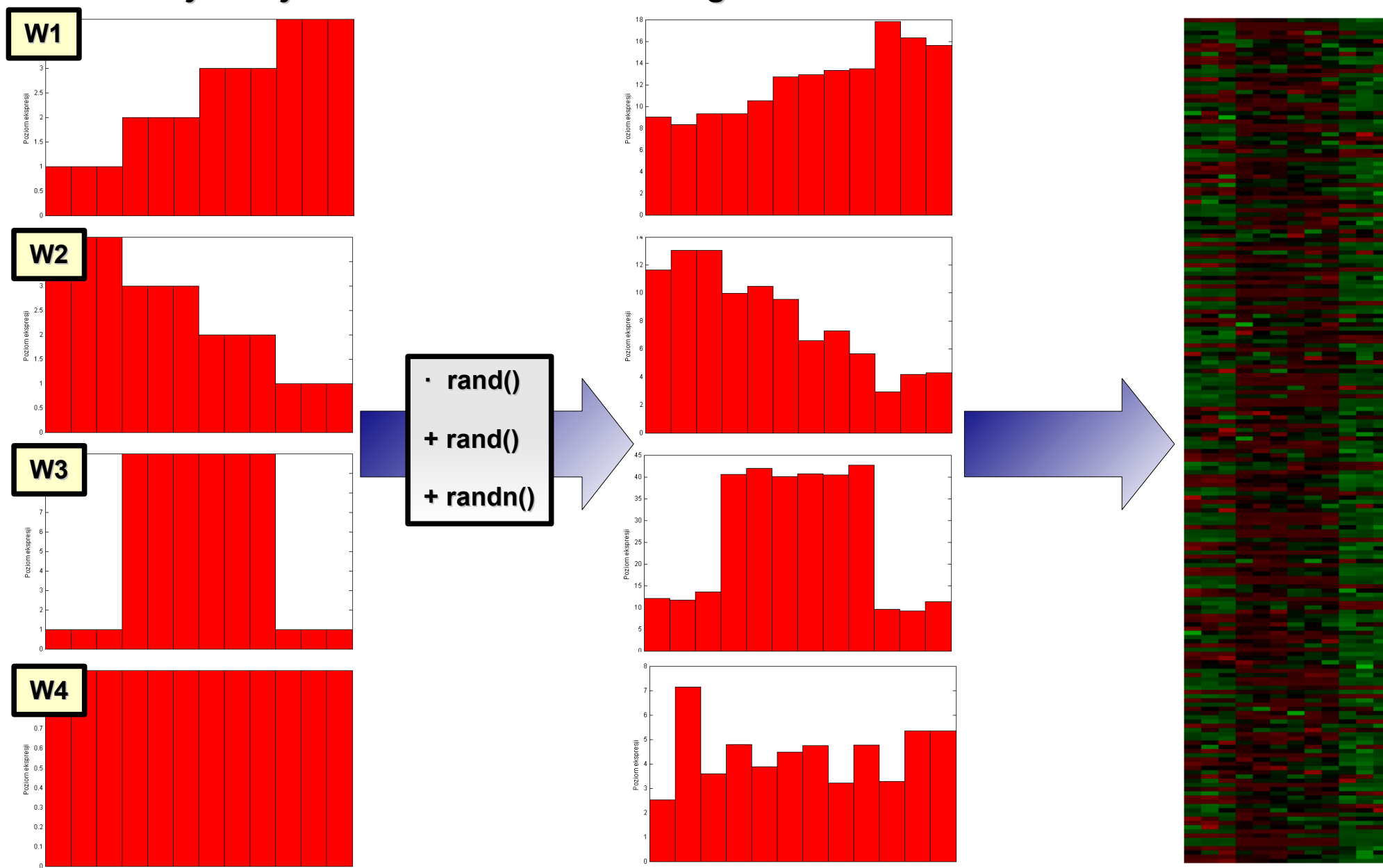
Inaczej: wybieramy geny o wartościach bezwzględnych w i -tej kolumnie macierzy T większych od zadanego progu.

- Wartości ekspresji j -tego genu są kombinacjami liniowymi wszystkich cech w przestrzeni składowych głównych. Wybieramy geny, w ekspresji których względny udział i -tej cechy z przestrzeni składowych głównych przekracza zadany próg.

Inaczej: wybieramy j -ty gen jeżeli w j -tym wierszu macierzy T stosunek modułu wartości i -tego elementu odniesiony do sumy modułów pozostałych elementów jest większy od pewnego progu.

Redukcja wymiarowości: zastosowania PCA (selekcja cech)

Przykład selekcji cech dla sztucznego zbioru danych złożonego z 200 „genów” o wartościach ekspresji wygenerowanych przez losowe przeskalowanie, dodanie składowej stałej oraz szumu Gaussowskiego do czterech wzorców:



Redukcja wymiarowości: zastosowania PCA (selekcja cech)

W wynikowym zbiorze danych dominują dwie pierwsze składowe główne, które odpowiadają wzorcom [W1 i W2] oraz [W3].

