ED – Laboratorium 4

# PCA

Iwo Błądek

## 1 Transformacje liniowe

Na transformacje liniowe można patrzeć jak na zmianę bazy przestrzeni liniowej. Wektor kolumnowy $\vec{x}$ wyrażony jest w „standardowym" układzie współrzędnych o bazie opartej na wektorach jednostkowych długości 1 (wersory). Po lewostronnym przemnożeniu przez pewną macierz $A$, otrzymujemy po prawej stronie oryginalny wektor przerzutowany do przestrzeni określonej kolumnami macierzy $A$.

$$\vec{x} = I \cdot \vec{x} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + x3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$A \cdot \vec{x} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x1 \begin{bmatrix} a \\ b \\ c \end{bmatrix} + x2 \begin{bmatrix} d \\ e \\ f \end{bmatrix} + x3 \begin{bmatrix} g \\ h \\ i \end{bmatrix} = \begin{bmatrix} ax_1 + dx_2 + gx_3 \\ bx_1 + ex_2 + hx_3 \\ cx_1 + fx_2 + ix_3 \end{bmatrix}$$

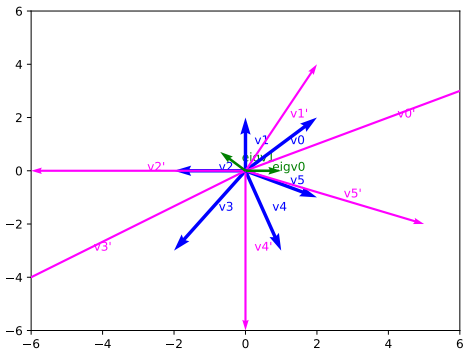
Jeżeli po prawej stronie macierzy $A$ byłaby macierz z $k$ kolumnami, to wynikowa macierz byłaby złożona z osobno przemnożonych kolumn:

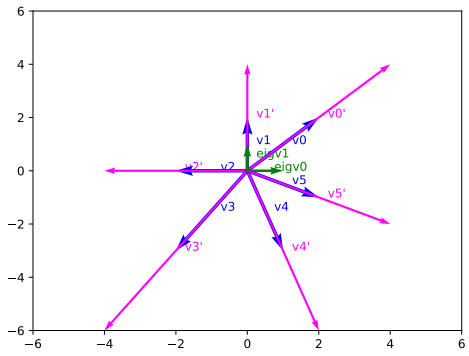
$$A \cdot \begin{bmatrix} \vec{x_1} & \cdots & \vec{x_k} \end{bmatrix} = \begin{bmatrix} A\vec{x_1} & \cdots & A\vec{x_k} \end{bmatrix}$$

## 2 Wektory i wartości własne macierzy

Każda macierz $A$ ma pewne wektory własne $\vec{k}$ (potencjalnie zespolone; jeżeli zawęzimy się do wektorów własnych rzeczywistych to macierz może takowych nie mieć, np. macierz obrotu o 90 stopni), które cechują się tym, że po pomnożeniu takiego wektora przez $A$ otrzymamy: $A \cdot \vec{k} = \lambda\vec{k}$. Innymi słowy, wynikiem mnożenia jest ten sam wektor przeskalowany o pewną stałą *lambda*. Długość i zwrot wektora $\vec{k}$ nie mają wpływu na tę własność (wynika to dość trywialnie ze wzoru), więc dla danej wartości własnej $\lambda$ jest nieskończona liczba wektorów własnych postaci $a \cdot \vec{k}$, gdzie $a$ dodatkowo skaluje pewien wybrany niezerowy „kanoniczny" (np. długości 1) wektor własny $\vec{k}$.

Na poniższych rysunkach na zielono zaznaczone są wektory własne przekształcenia (realizowanego przez macierz) z wektorów niebieskich na różowe. Im bardziej kierunek wektora jest zgodny z którymś z wektorów własnych macierzy przekształcenia, tym bliżej oryginalnej postaci wypada wektor po przekształceniu (jeżeli $\lambda < 0$, to wtedy wylądowałby po drugiej stronie jako odbicie przez środek układu współrzędnych).




**Ćwiczenie 2.1:** Czy potrafisz na podstawie powyższych wykresów podać wartości własne tych macierzy przekształcających?
**Ćwiczenie 2.2:** Co można powiedzieć o macierzy przekształcenia z drugiego wykresu? Czy potrafisz bez obliczeń podać jej postać? Czy istnieje dla niej niezerowy wektor, który nie byłby jej wektorem własnym?

**Ćwiczenie 2.3:** Oblicz wartości własne i wektory własne macierzy $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$.

**Ćwiczenie 2.4:** Oblicz wartości własne i wektory własne macierzy $\begin{bmatrix} -1 & 2 \\ 2 & 1 \end{bmatrix}$.

# 3 Rozkład EVD

Jeżeli znamy wartości i wektory własne macierzy to możemy ją zdekomponować na następujący iloczyn:

$$A = KLK^{-1}$$

gdzie $K$ zawiera w kolumnach ortogonalne wektory własne, a $L$ jest macierzą diagonalną i ma odpowiednie wartości własne na przekątnej. Wynika to z definicji wektorów własnych: $AK = KL$.

Jeżeli macierz $A$ jest symetryczna, to $K$ jest *ortogonalna* i zachodzi $K^{-1} = K^T$, co znacznie upraszcza obliczanie odwrotności macierzy. Ten fakt zostanie wykorzystany przy obliczaniu PCA.

**Ćwiczenie 3.1:** Macierz $A$ ma wartości własne $\lambda_1 = 4$, $\lambda_2 = 1$ i odpowiadające im wektory własne $\begin{bmatrix} 1 & 2 \end{bmatrix}$ i $\begin{bmatrix} -1 & 1 \end{bmatrix}$. Podaj macierze $K$ i $L$, oraz rozkład EVD macierzy $A$.

# 4 PCA

Analiza składowych głównych, *PCA* (ang. *Principal Component Analysis*), przekształca dane do układu współrzędnych, w którym wariancje zmiennych (atrybutów) maleją wraz z kolejnymi wymiarami i są kolejno maksymalizowane (tzn. nie ma przekształcenia liniowego dającego większą pierwszą wariancję, itd. dla pozostałych). PCA działa w następujących krokach, gdzie $X$ to wycentrowana macierz oryginalnych danych (wiersze – przypadki, kolumny – atrybuty):

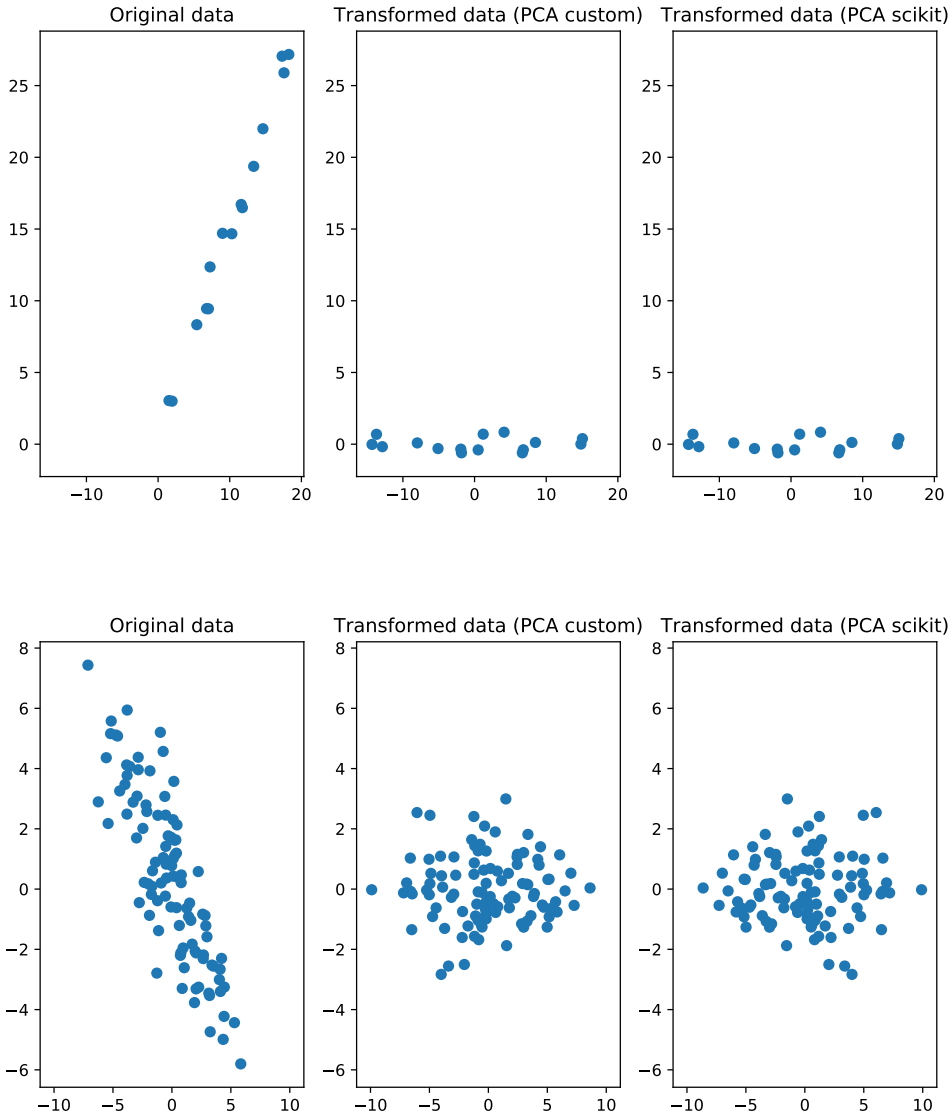
1. Obliczenie macierzy kowariancji: $S_x = X^T X$ (ułamek $\frac{1}{m}$ może zostać pominięty). Jest to macierz symetryczna.
2. Zastosowanie EVD na otrzymanej macierzy kowariancji: $S_x = KLK^{-1} = KLK^T$. Celem jest otrzymanie macierzy wektorów własnych $K$. Jest ona ortogonalna, co pozwoliło nam zamienić odwrotność na transpozycję. Ortogonalność $K$ wynika tutaj z symetryczności $S_x$.
3. Otrzymanie nowych zmiennych: $Y = XK$.

# 5 Zadanie domowe (3 punkty)

**Zadanie 5.1:** Zaimplementuj algorytm PCA korzystając z szablonu na stronie i wypełniając w nim ciało funkcji `pca_manual`. Uzupełnij w niej również instrukcje print tak by wypisywały odpowiednie informacje.

**Zadanie 5.2:** Uzupełnij w szablonie funkcję `pca_sklearn` obliczającą PCA przy użyciu biblioteki scikit-learn (`http://scikit-learn.org/stable/`). Klasa odpowiedzialna w tej bibliotece za PCA to `sklearn.decomposition.PCA`. Zapoznaj się z jej dokumentacją w internecie.

Poniżej przedstawione są przykładowe wyniki, jakie można uzyskać. Dla drugiego zbioru danych można zauważyć, że wykresy dla implementacji PCA są różne – wynika to z wybrania innych wektorów własnych w macierzy $K$ (w szczególności: wektorów przeciwnych).




**Zadanie 5.3: (Opcjonalne)** Wektory własne w macierzy $K$ są wzajemnie ortogonalne i mają długość równą 1. Oznacza to, że można przemnożyć w $K$ dowolną kolumnę (wektor własny) przez $-1$ i nie zmienić powyższych własności. Praktyczna konsekwencja jest taka, że różne implementacje PCA mogą dać różne wyniki zależnie od wybranych wektorów własnych (jednak wariancje zmiennych zawsze będą takie same).

Dodaj w szablonie kod, który sprawdzi, czy Twoja implementacja daje taki sam (z dokładnością do pewnego $\epsilon$) wynik jak implementacja w scikit-learn dla którejś z kombinacji znaków wektorów własnych z $K$.

Zmodyfikuj również kod w taki sposób, by na wykresie przedstawiony był wynik Twojej implementacji najbardziej zbliżony do tego ze scikit-learn.