Analiza głównych składowych (PCA)

w analizie danych spektroskopowych

mgr Damian K. Chlebda

Uniwersytet Jagielloński w Krakowie
[k] zespół kinetyki reakcji heterogenicznych
pap deg lab

Podstawy > Definicja

Analiza głównych składowych
(ang. *Principal Component Analysis*, PCA)

najbardziej wszechstronne narzędzie w danych analizy

celem jest wyznaczenie nowych nieskorelowanych zmiennych (głównych składowych), które będą miały największą możliwą wariancję

Metoda wyznaczenia korelacji pomiędzy zestawami danych spektralnych; opisuje dane, nie przez długość fali lecz używając mniejszej liczby „ukrytych” zmiennych zwanych głównymi składowymi (ang. *principal components* (PCs)).

Ogólnie działanie PCA opiera się na założeniu, że:

DANE = INFORMACJA + SZUM

W konsekwencji, informacja definiowana jest przez pierwsze PCs, a szum jako pozostałość.

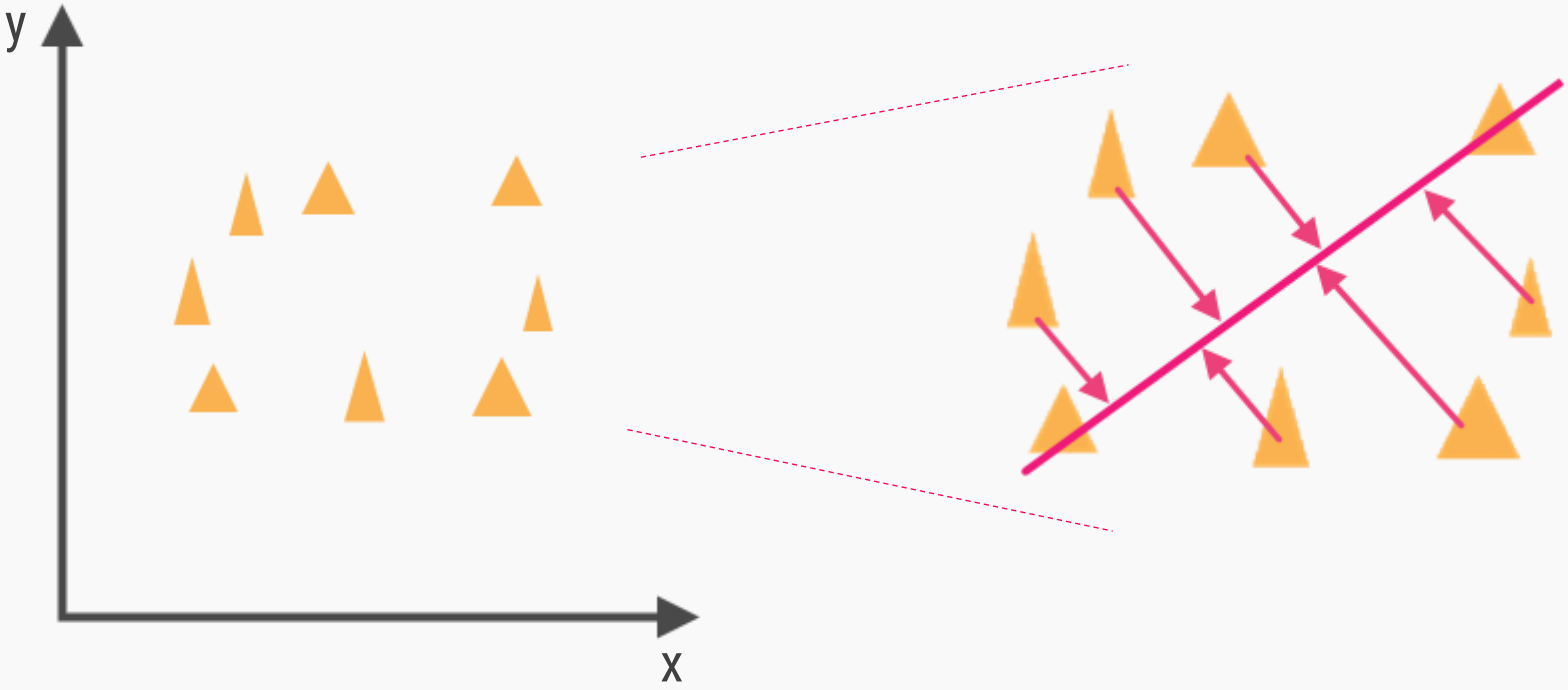
Suma PCs i szumu musi dawać dane oryginalne / wejściowe.

Możliwości analizy PCA

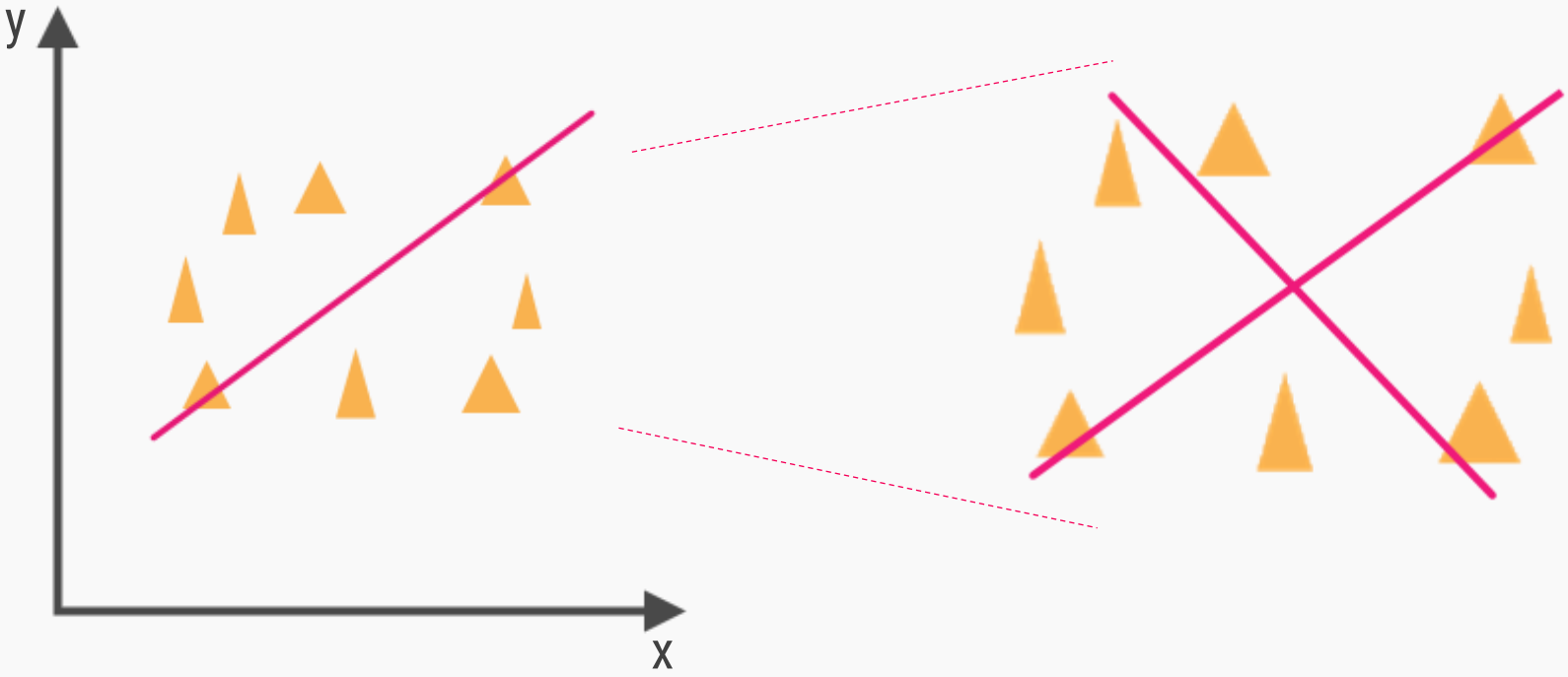
- pozwala zwizualizować wszystkie informacje zawarte w zestawie danych;
- pomaga dowiedzieć się, pod jakim względem jedna próbka różni się od innych;
- które zmienne przyczyniają się najbardziej do tych różnic;
- czy te zmienne przyczyniają się w ten sam sposób (są skorelowane) lub niezależnie od siebie,
- pozwala wykryć wzory w próbkach;
- oddziela dużą ilość użytecznych informacji od szumu;
- pozwala na redukcję danych.

Ważnym jest poznanie działania PCA, gdyż jest to użyteczna metoda analizy i stanowi podstawę paru metod klasyfikacji (SIMCA) i regresji (PLS/PCR).

1. Konstrukcja czynników głównych, mając na uwadze, że powinny one jak najlepiej opisać wariancję danych.

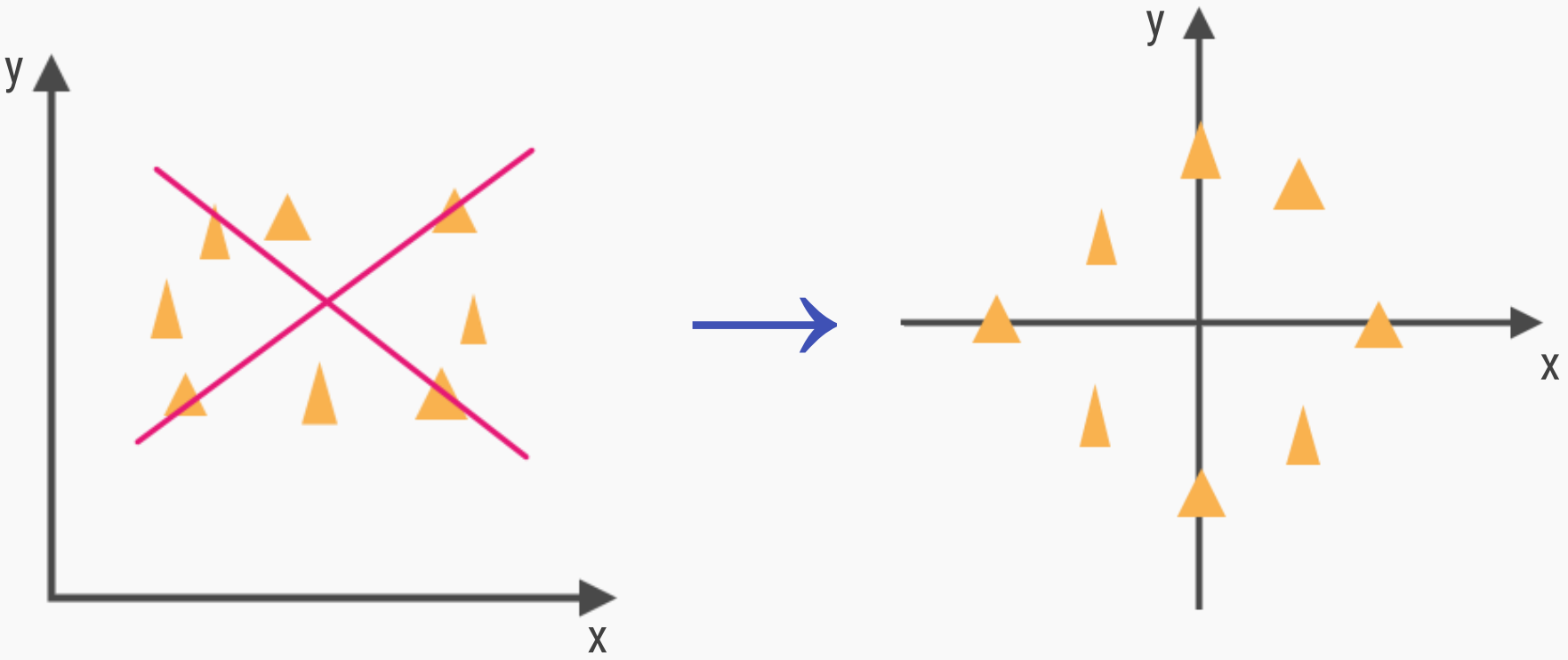


2. Po wyznaczeniu pierwszej zmiennej wyznaczamy drugą, tak by była ortogonalna do pierwszej, i wyjaśniała możliwie dużo pozostałej zmienności.



* W przypadku danych wielowymiarowych (3D i więcej) kolejną zmienną wybieramy tak, by była ortogonalna do dwóch pierwszych itd..

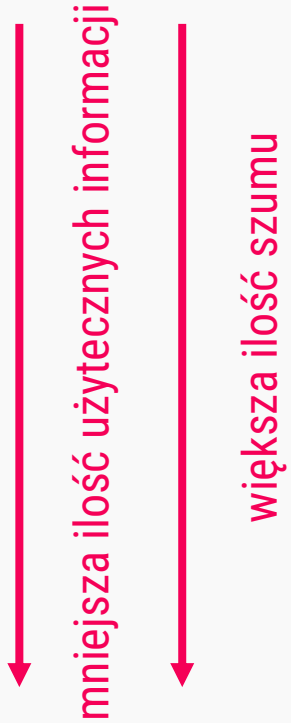
3. Tak uzyskany zbiór wektorów tworzy bazę ortogonalną w przestrzeni cech. Pierwsze współrzędne wyjaśniają większość zmienności danych.



- Celem metody składowych głównych jest więc znalezienie transformacji układu współrzędnych, która lepiej opisze zmienność pomiędzy obserwacjami
- W nowym układzie współrzędnych, odległości euklidesowe pomiędzy obiektami są zachowane (tzn. są równe odległościom w przestrzeni oryginalnych zmiennych).

Główne składowe (PC) są liniowymi funkcjami oryginalnych zmiennych i zawierają kolejno, w porządku malejącym, informacje dotyczące struktury zmienności danych.

główna składowa	
1	wyjaśnia największą część zmienność danych i zawiera najwięcej informacji
2	jest ortogonalna (prostopadła) do pierwszej; reprezentuje największą zmienność wokół pierwszej głównej składowej
3	jest ortogonalna do pierwszej i drugiej; opisuje największą zmienność nie wyjaśnianą przez te składowe
...	



Informacje wyjściowe

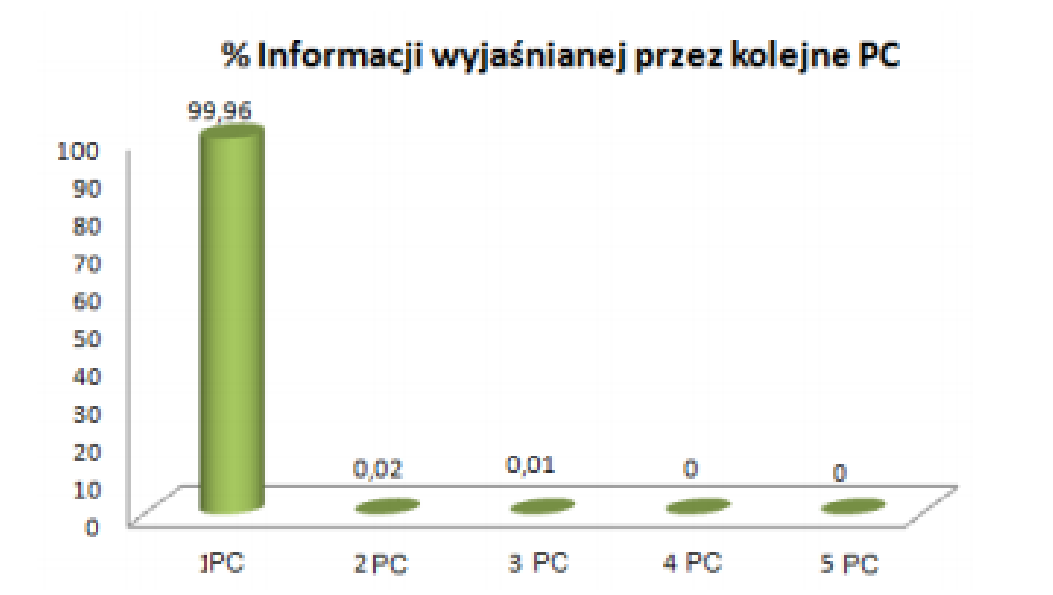
Informacje otrzymywane po przeprowadzeniu analizy głównych składowych można podzielić na dwie grupy:

Wykres wag / *Scores plot* → przedstawia mapę próbek

Wykres ładunków / *Loadings plot* → przedstawia mapę zmiennych czynnikowych

Wyniki analizy > Wykres wariancji / osypiska / piargowy

Kumulacyjny procent wariancji danych opisanej przez pierwsze x czynników głównych.

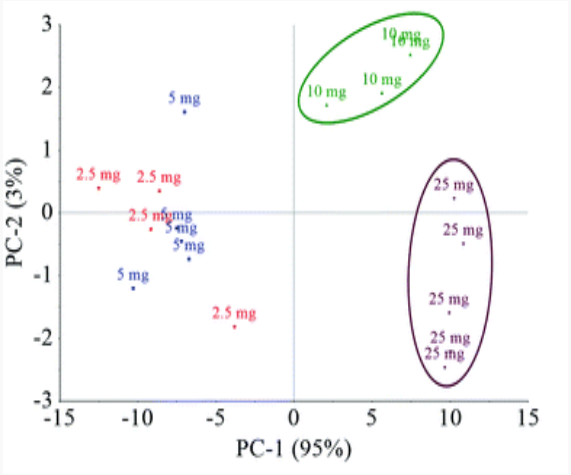


Analizując dane z PCA najpierw jest analizowany wykres wartości własnych (*eigenvalue*) w celu określenia liczby składowych, które odzwierciedlają różnice między danymi.

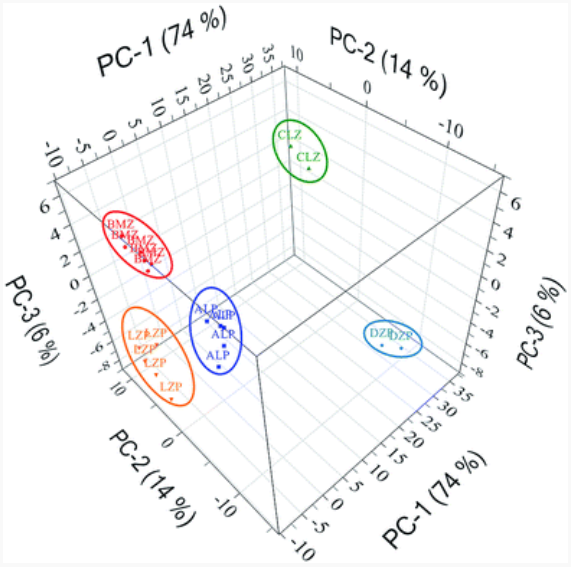
Ile składowych wybrać do dalszej analizy?

Wyniki analizy > PC vs PC

2D:



3D:



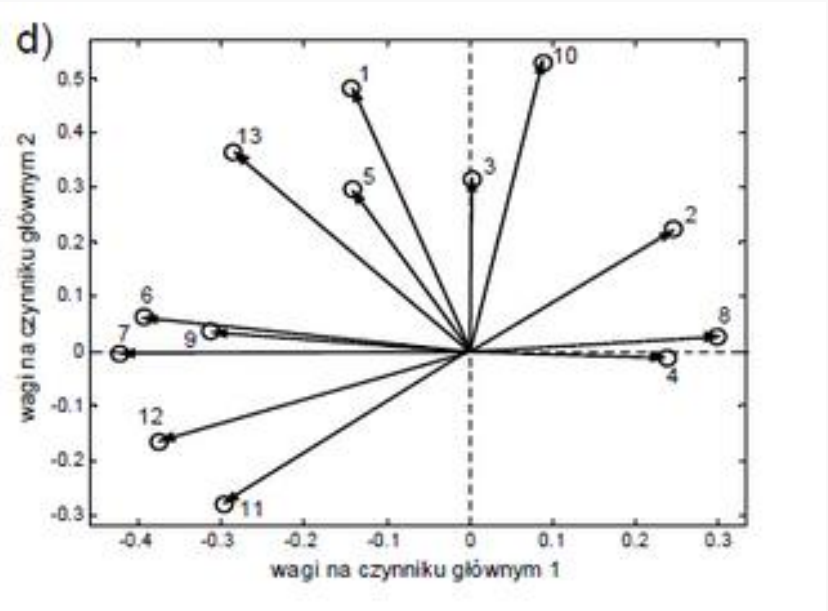
przedstawienie położenia próbek w nowym układzie współrzędnych zdefiniowanym przez pierwsze czynniki główne

- pozwalają ukazać niehomogeniczną strukturę danych
- są źródłem informacji o tendencji danych do grupowania
- i/lub o próbkach, które znacząco różnią się od pozostałych (tak zwanych obiektów odległych)
- jako miarę podobieństwa pomiędzy próbkami wykorzystuje się odległość euklidesową

↓ im mniejsze tym bardziej podobne ↓

Wyniki analizy > Projekcja wag

dokonuje się projekcji wag na płaszczyzny zdefiniowane parami czynników głównych



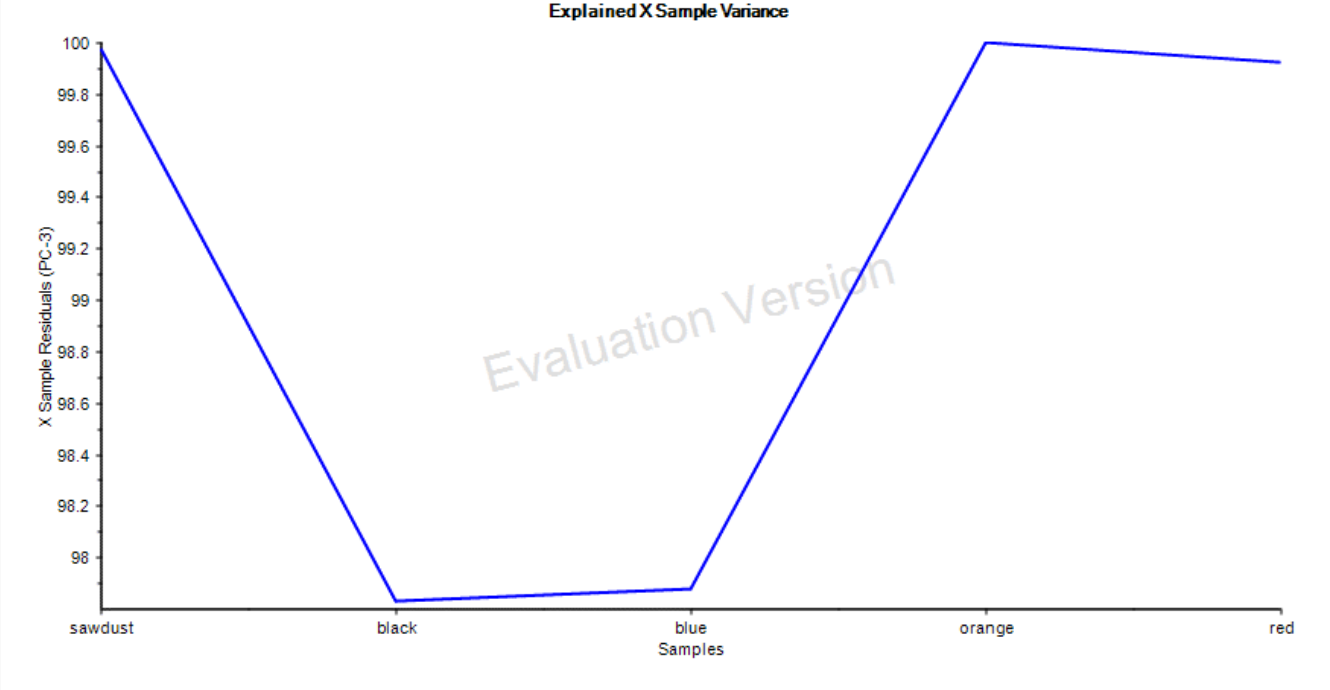
- pozwala zbadać, które z parametrów są do siebie podobne, a które różnicują próbki
- wzajemne podobieństwa określa się na podstawie kąta, jaki tworzą pomiędzy sobą dwa wektory wag o początku w punkcie [0 0] i końcach zdefiniowanych przez odpowiednie wartości wag zmiennych na rozważanych projekcjach

- bliski 0° wówczas są one silnie dodatnio skorelowane;
- bliski 180° to parametry są silnie skorelowane, ale przeciwnie;
- bliski 90° - dwa parametry są niezależne (ortogonalne).

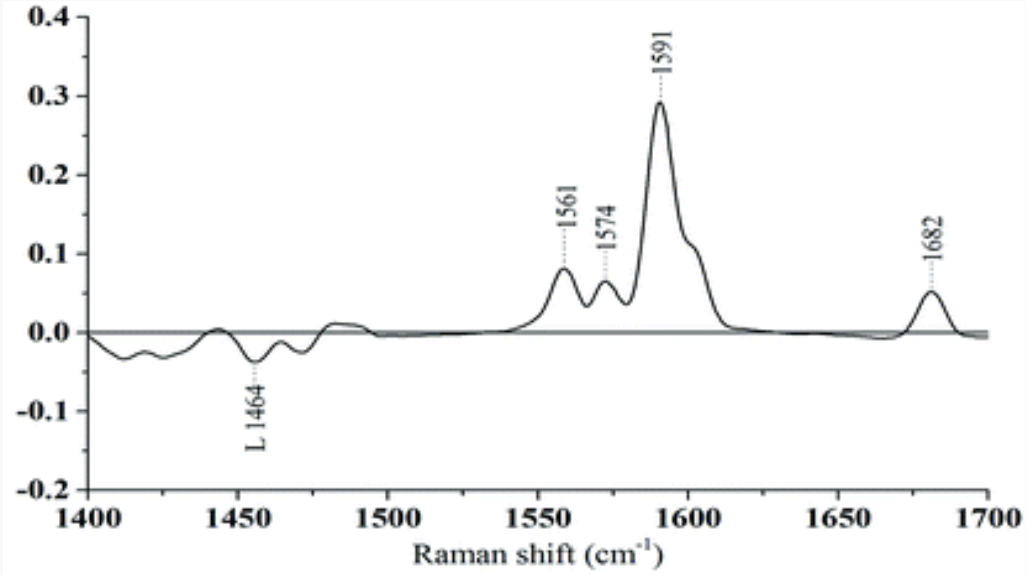
przedstawiają one jedynie pewną część wariancji danych!!!

Wyniki analizy > PC vs próbka

- Pozwala zobaczyć, które próbki opisywane są najlepiej przez daną PC



Wyniki analizy > *Loadings* vs zmienna (*variables*)



Loadings (for PC-1) for four different doses of Kern Pharma Diazepam in the 1400 to 1700 cm−1 spectral range

- Patrząc na wykres PC vs próbka -> które próbki opisuje dana składowa
- Opis pasm charakteryzujących daną grupę (różniących je od innych)

- *Loadings* dostarcza informacji o strukturze PC powiązaną z daną zmienną (spektralną)
- *Loadings* dla danych spektralnych lepiej jest interpretować obserwując je oddzielnie na widmach, po jednym na raz, w funkcji długości fali

Wyniki analizy > *Inne wykresy*

- **Wykres wpływu (*the influence plot*)** przedstawia w jednym wykresie, jak dobrze każda próbka jest odwzorowana (z odległości X-rezydualnych) i jak daleko każda próbka znajduje się od ogólnego środka modelu. Wartości odległości rezydualnej powinny być zbliżone do zera dla dobrze opisanych próbek. Model przedstawiany jest w skali od zera do jednego, gdzie 1 opisuje próbki o skrajnych właściwościach.

- **Wykresy próbek i zmiennych wykresy (*sample and variable residual plots*)** opisują znaczenie każdej zmiennej i jak dobrze próbka jest reprezentowana w tym modelu.

Wyniki analizy > *Podsumowanie*

- PCA prowadzi do redukcji zmiennych / danych.
- Analiza wykresów ładunków i wartości pozwala znaleźć i opisać regiony widma związanych z określonymi grupami próbek.
- Analiza pozwala w szczegółowy sposób opisać analizowaną pulę próbek; dokonać klasyfikacji i wyodrębnić cechy próbek będących podstawą do dyskryminacji

Planowanie analizy > podejście chemometryczne

Wybór metody statystycznej przy analizie danych spektralnych wymaga określenia celu analizy. Istnieją różne metody i programy statystyczne, jednak w każdym przypadku należy określić precyzję, dokładność, heterogeniczność i główny cel porównania tak, aby wybrać właściwe narzędzie i zminimalizować błąd analizy.

Odpowiedź na pytanie:

a) Jaką techniką rejestrowane były widma? → spodziewana precyzja, rozdzielczość, dokładność

b) Jaki jest skład matrycy? → heterogeniczność materiału również gra istotną rolę w analizie podobieństwa

c) Jaki jest cel analizy? → np. porównanie próbek by określić ich podobieństwo/grupowanie; porównanie mające na celu ustalenie wspólnego pochodzenia; porównanie mające na celu wskazanie istotnych różnic przebiegu widma

d) Jak dużo powtórzeń pomiaru (widma) przypada na próbkę?

Na przygotowanie danych do analizy PCA składa się:

1. Korekta linii bazowej (lub SVN, MSC)
2. Centrowanie
3. Obliczenie 1 pochodnej
4. Normalizacja
5. Autoskalowanie danych

Ma na celu wyeliminowanie stałych elementów, które nic nie wnoszą do zróżnicowania danych.

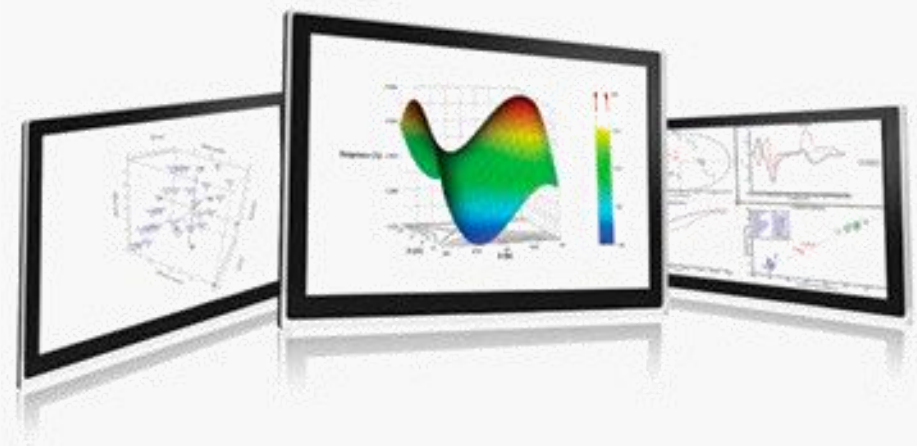
Ma na celu usunięcie efektu związanego z różną ilością próbki użytej w eksperymencie.

Stosowane, gdy zmienne eksperymentalne są w różnych jednostkach i/lub ich odchylenia standardowe znacznie się różnią; łączy w sobie dwie operacje: centrowania i standaryzacji.

Strona praktyczna > Programy obliczeniowe

Ważne jest, aby wybrać oprogramowanie, które jest zarówno odpowiednie do Twoich potrzeb.

- Unscrambler X,
- Origin,
- Statistica,
- WinISI lub R,
- Matlab,
- GRAMS IQ™ Spectroscopy Software,
- xlstat itd.



Więcej na: *http://www.wiley.com/legacy/wileychi/chemometrics/software.html*

Strona praktyczna > Rady

Przy interpretacji danych wyjściowych modelu PCA, weź pod uwagę następujące wytyczne:

- Zawsze najpierw popatrz na wykres wariancji (*the explained variance plot*), aby zobaczyć ile PCs jest wymaganych aby rozsądnie opisać część zmienności. Im mniej tym zazwyczaj lepiej. Jeśli do opisu < 50 % danych potrzeba dużo PCs, dane mogą zawierać w większości szum.

- Po zdecydowaniu ile PCs użyć, zobacz wykres wag (*scores plot*) PC1 vs PC2 (i kombinacje innych PCs których zdecydowałeś się użyć, i poszukaj tam zależności. Zależności te mogą ujawnić istnienie mniejszych struktur danych dotyczących różnych zjawisk zachodzących w danych.

Strona praktyczna > Rady

Przy interpretacji danych wyjściowych modelu PCA, weź pod uwagę następujące wytyczne:

- *Loadings plots* można użyć do interpretacji, które zmienne najbardziej przyczyniają się do grupowania próbek (jeśli istnieją). Pomaga to uzyskać lepszy obraz tego, co dzieje się w twoim zestawie danych, a także ujawnia wszystkie zmienne, które współdziałają ze sobą.

- Użyj wykres wpływu (*the influence plot*), aby ustalić, czy wszystkie próbki zostały opisane dobrze przez model a wartości wag do ustalenia czy obecne są skrajne próbki.

- Proste modele zawierają najmniejszą liczbę PCs i są najłatwiejsze do interpretacji, dlatego zawsze należy stosować zasadę **KISS** (Keep It Simple, Stupid)

Ograniczenia metody PCA

- Analiza głównych składowych nie działa, gdy dysponujesz mniejszą ilością widm, niż jest składników w próbce.
- Wszystkie widma muszą mieć wspólną oś odciętych.
- Należy zadbać o dobrą liczebność próby

PCA i klasyfikacja

wyraźne i łatwe do definiowania grupy w zbiorze danych

opracowanie oddzielnych modeli PCA dla tych grup

Soft Independent Modeling of Class Analogy (SIMCA).