Analiza danych

prof. UAM dr hab. Tomasz Górecki

tomasz.gorecki@amu.edu.pl
http://drizzt.home.amu.edu.pl

Zakład Statystyki Matematycznej i Analizy Danych
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Redukcja wymiaru – wprowadzenie

Dlaczego warto to robić:

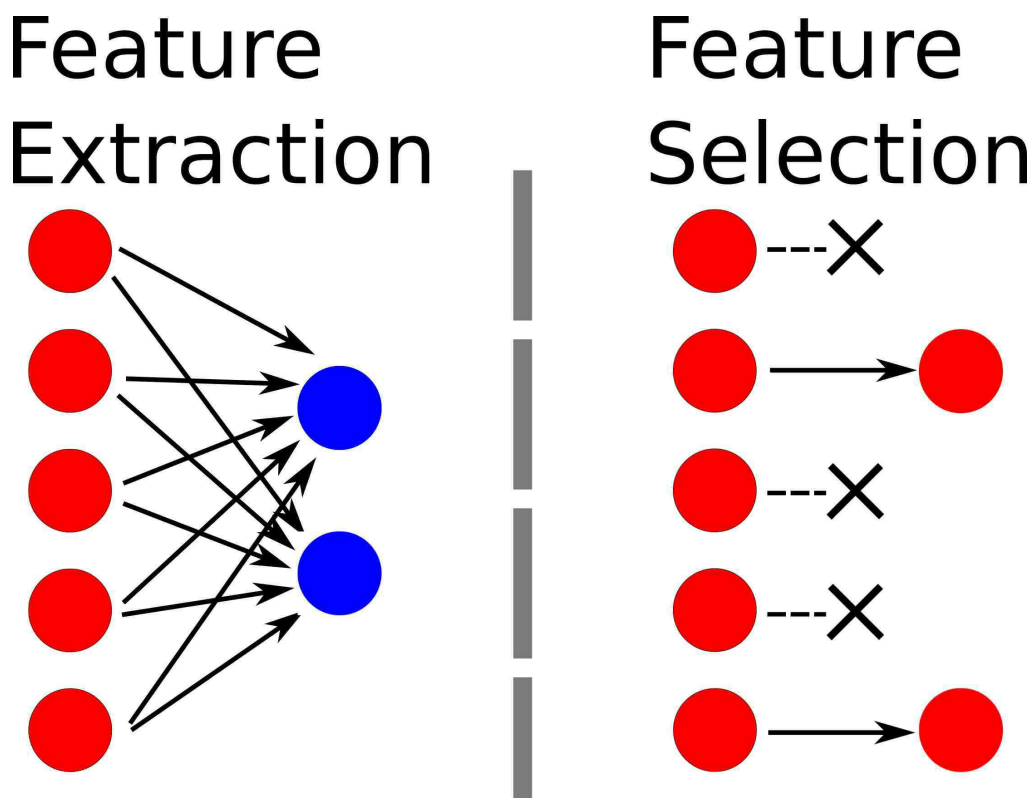
- Jakość modelu – redukcja liczby cech może prowadzić do poprawy jakości modelu.
- Przeuczenie – nadmiar cech może prowadzić do przeuczenia.
- XAI (*ang. eXplainable AI*) – im więcej zmiennych w modelu tym trudniej będzie wyjaśnić biznesowi, jak działa model.
- Czas uczenia – mniej danych oznacza, że modele uczą się szybciej.
- Wdrożenie i utrzymanie – im mniej cech tym ten proces będzie prostszy.

Redukcja wymiaru – wprowadzenie

Redukcja wymiaru (*ang. dimensionality reduction, dimension reduction*) – proces zmniejszania liczby zmiennych branych pod uwagę podczas analizy, w taki sposób aby zachować jak najwięcej istotnych informacji. Redukcja wymiaru może polegać na:

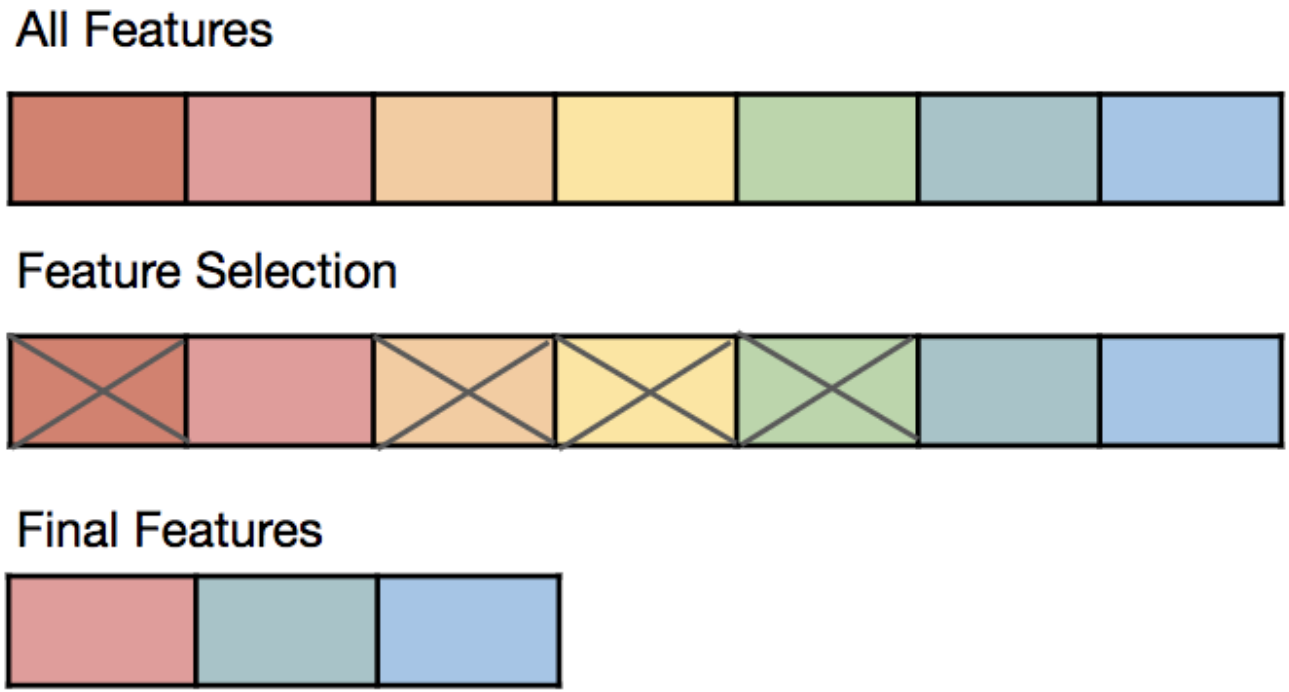
1. Selekcji cech (*ang. feature selection*) – ograniczeniu zbioru zmiennych według pewnych reguł (np. cechy nadmiernie skorelowane ze sobą, cechy nieistotne statystycznie)
2. Ekstrakcji cech (*ang. feature extraction*) – tworzeniu nowych cech pochodnych z początkowego zestawu danych celem uzyskania mniejszego zbioru zmiennych.

Redukcja wymiaru – wprowadzenie

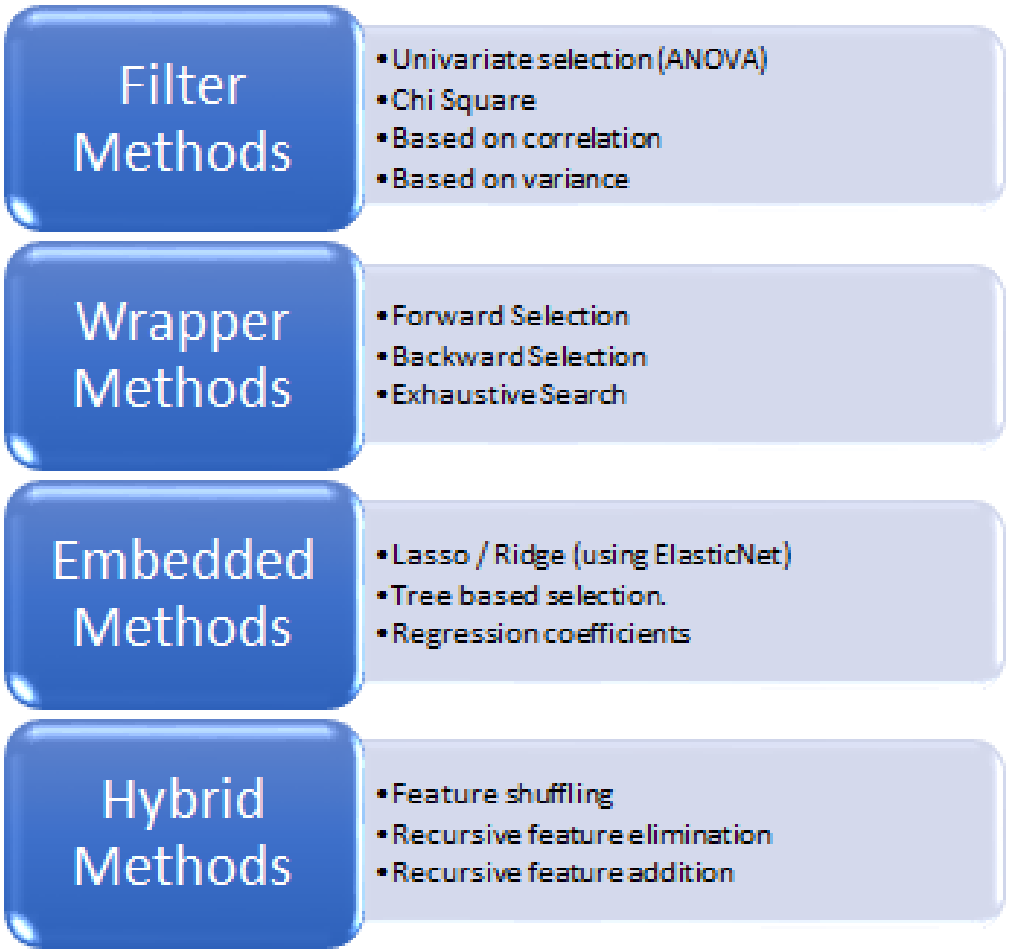


Wybór zmiennych – wprowadzenie

Wybór zmiennych (*ang. feature selection*) – proces, w którym wybieramy cechy, które mają największy wpływ na zmienną przewidywaną. Jest to zatem proces wybierania podzbioru zmiennych z wszystkich dostępnych danych, aby uzyskać jak najlepszy model.

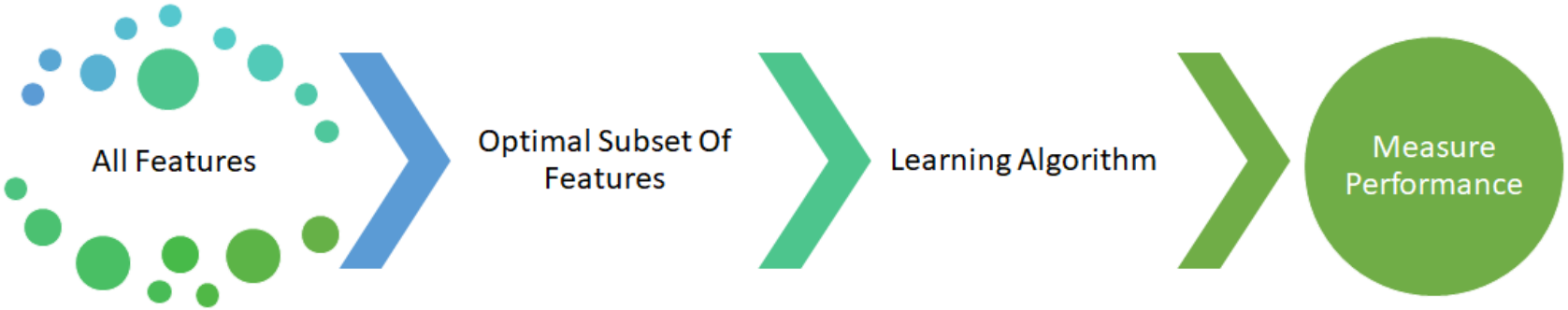


Wybór zmiennych – wprowadzenie



Metody oparte na filtrach

Ostateczną listę cech otrzymujemy na podstawie przyjętej metody filtrowania. Metody te opierają się na cechach danych i nie używają algorytmów uczenia maszynowego. Zwykle są one szybsze kosztem niższej wydajności względem pozostałych metod.



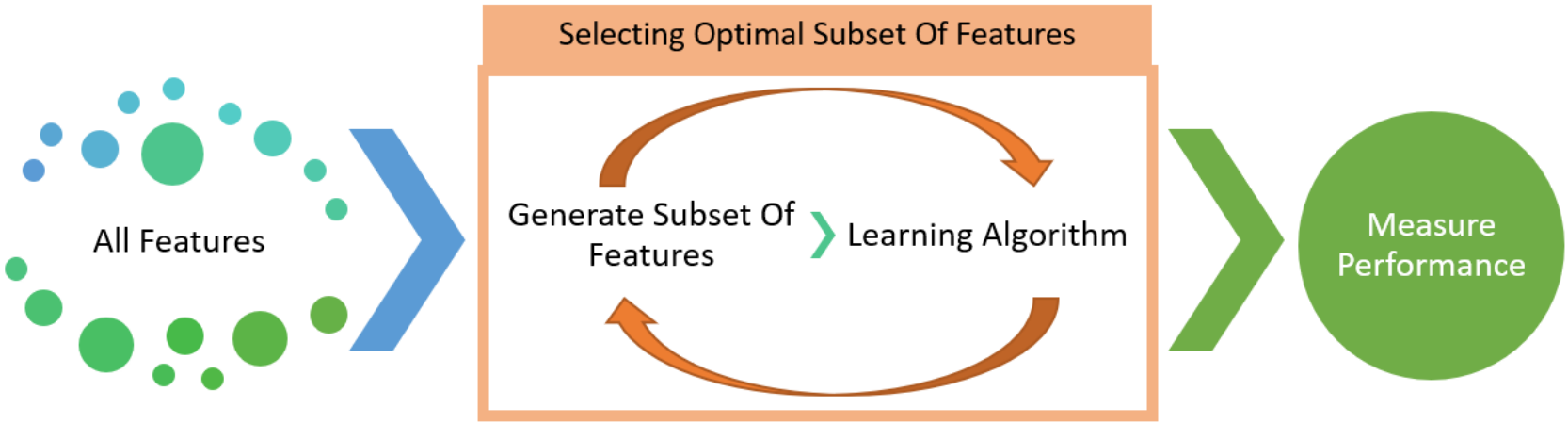
Metody oparte na filtrach

Podstawowe metody:

- Próg wariancji (*ang. variance threshold*) – usuwamy cechy o stałej wariancji.
- Współczynnik korelacji PEARSONA – dobre zmienne powinny być mocno skorelowane ze zmienną docelową.
- Statystyka χ^2 – testujemy związek między cechami w zbiorze danych, a zmienną docelową. Obliczamy wartość statystyki χ^2 między każdą cechą a celem i wybieramy pożądaną liczbę cech z najlepszymi wynikami.

Metody oparte na wrapperach

Są to metody, które traktują wybór zmiennych jako problem wyszukania. Proces wyboru cech opiera się na określonym algorytmie uczenia maszynowego, który staramy się dopasować do danego zbioru danych szukając optymalnego zestawu cech. Na podstawie wniosków wyciągniętych z poprzedniego modelu decydujemy o dodaniu lub usunięciu cech. Metody te są zwykle bardzo kosztowne obliczeniowo.



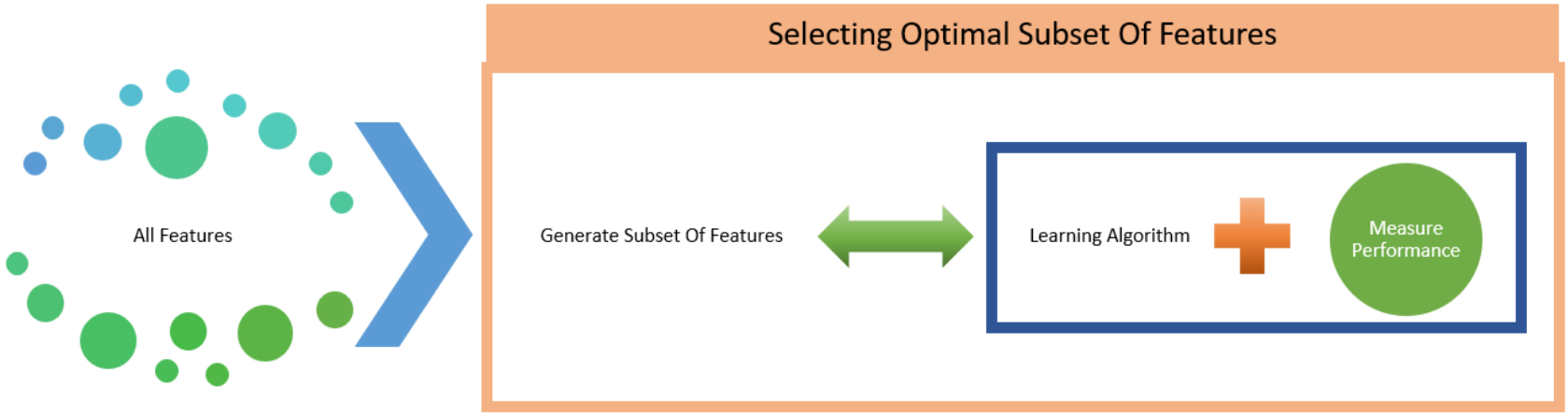
Metody oparte na wrapperach

Podstawowe metody:

- Przeszukiwanie w przód (*ang. forward feature selection*)
- Przeszukiwanie w tył (*ang. backward feature elimination*)
- Przeszukiwanie wyczerpujące (*ang. exhaustive feature selection*)

Metody oparte na osadzeniach (*ang. embedded methods*)

Metody osadzone wykorzystują algorytmy, które mają wbudowane metody wyboru cech. Ogólnie są mniej wymagające obliczeniowe niż metody oparte na wrapperach. Często dają najlepsze wyniki.



Metody oparte na osadzeniach (*ang. embedded methods*)

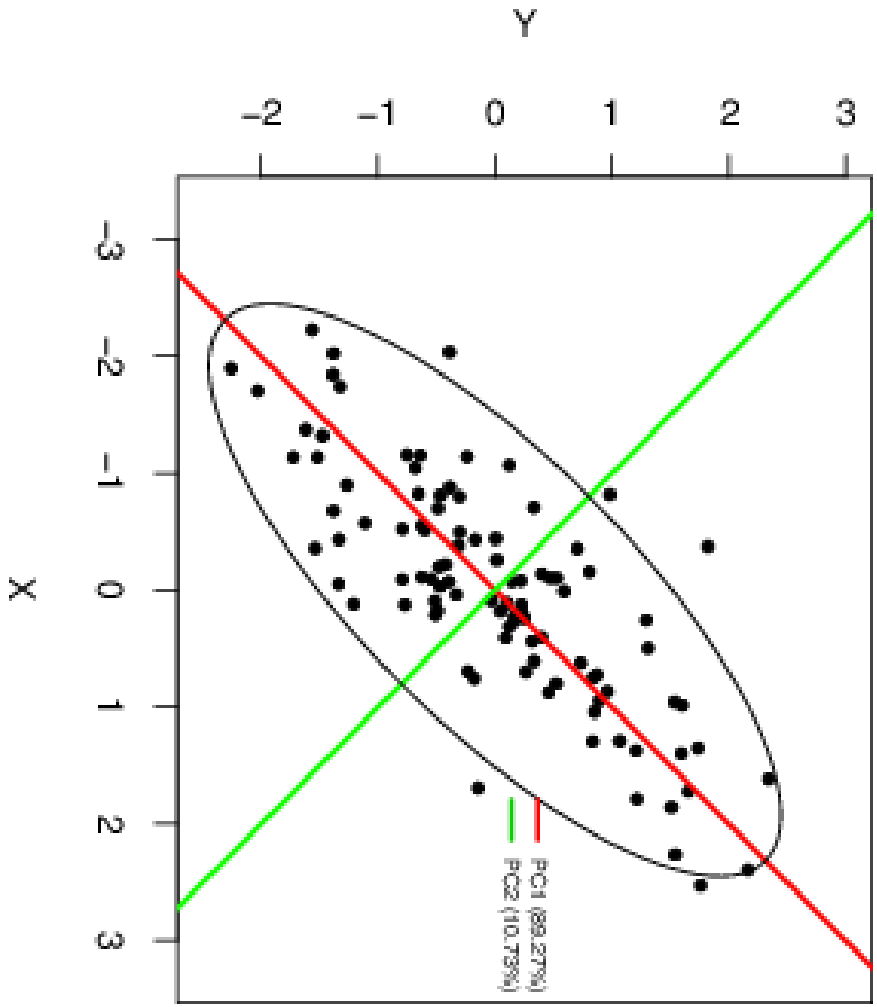
Podstawowe metody:

- LASSO
- Las losowy

Idea

Analiza składowych głównych (*ang. principal components analysis* – PCA), zwana również dekompozycją według wartości osobliwych (SVD) lub dekompozycją spektralną, jest popularną techniką redukcji wymiarowości danych (liczby cech). Jest to metoda nieparametryczna, a co za tym idzie nie wymaga żadnych założeń, co do rozkładów badanych danych. W metodzie tej chcemy zastąpić zbiór skorelowanych cech (jeśli zmienne nie są skorelowane, PCA nie daje redukcji danych) przez małą liczbę nieskorelowanych tzw. składowych głównych, które razem mogą wyjaśnić prawie całą zmienność danych. Pierwsza składowa wyjaśnia najwięcej zmienności (składowe są kombinacjami liniowymi wejściowych zmiennych). Druga składowa wybierana jest w taki sposób, aby nie była skorelowana z pierwszą i wyjaśniała jak najwięcej z pozostałej zmienności.

Idea



Historia

Składowe główne zostały po raz pierwszy zaproponowane przez PEARSONA (1901), a rozwinięte przez HOTELLINGA (1933, 1936).

Karl Pearson (1857-1936)

Harold Hotelling (1895-1973)

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24:417–441 and 498–520.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28:321–377.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2(11):559–572.

Ładunki i wyniki

W rezultacie otrzymujemy tyle składowych (są wzajemnie ortogonalne) ile było zmiennych oryginalnych, ale najczęściej jedynie kilka z nich wyjaśnia prawie całą zmienność danych. Jako wynik otrzymujemy najczęściej dwa elementy: ładunki (*ang. loadings*) oraz wyniki (*ang. scores*). Ładunki to współczynniki pokazujące wkład poszczególnych zmiennych oryginalnych w tworzeniu składowych głównych. Im wartość bezwzględna z ładunku większa tym zmienna ma większy wkład w budowę składowej głównej. Wyniki są współrzędnymi obserwacji w nowym układzie współrzędnych utworzonym przez składowe główne, to one najczęściej podlegają wizualizacji.

Liczba składowych

Jeśli chcemy zredukować wymiar danych musimy się zastanowić ile składowych wybrać do dalszej analizy. Najczęściej decyzję tę podejmuje się bazując na wykresie osypiska (piargowym). Jako optymalną liczbę czynników wybieramy tę, gdzie wykres się znacząco spłaszcza. Drugim popularnym kryterium jest ustalenie pewnego poziomu wariancji jaki muszą wyjaśnić składowe główne (najczęściej 90%). Kryterium osypiska prowadzi niekiedy do odrzucenia zbyt wielu czynników, ale w typowych sytuacjach (niezbyt dużo czynników i sporo obserwacji) radzi sobie całkiem dobrze.

Wizualizacja

Na koniec możemy zwizualizować nowe dane na jednym wykresie, na którym jako punkty będą przedstawione poszczególne obserwacje w układzie dwóch pierwszych składowych głównych, natomiast wektory oznaczać będą cechy. Kierunek wektorów pokazuje wpływ tych cech odpowiednio na pierwszą i drugą składową. Kąt przecięcia strzałek jest proporcjonalny do zależności pomiędzy cechami (dokładnie iloczyn skalarny odpowiednich wektorów wyznacza korelację), a ich długość odzwierciedla odchylenie standardowe. Tego typu wykres nazywa się biplotem.

Niezmienniczość

Składowe główne nie są niezmiennicze względem zmiany skali zmiennych oryginalnych. Oznacza to, że przeskalowanie danych zmienia wyniki analizy metodą PCA. Z tego względu składowe główne uzyskane z macierzy kowariancji oraz korelacji różnią się. W przypadku dużych różnic w wariancjach lub cech mierzonych na różnych skalach należy wpierw przeskalować dane (działać na macierzy korelacji).

Idea

W wielu dziedzinach nauki (zwłaszcza psychologii i naukach społecznych) nie jest możliwe zmierzenie wszystkich zmiennych bezpośrednio. W takim przypadku zbieramy informacje o zmiennych stowarzyszonych, które w pewien sposób wpływają na interesujące nas zjawisko. Przykładowo badając poziom inteligencji badamy ją za pomocą licznych testów. Zmienne nieobserwowalne nazywane są zmiennymi utajonymi (*ang. latent*). W takiej sytuacji używa się analizy czynnikowej (*ang. factor analysis (FA)*) do zidentyfikowania tych ukrytych zmiennych (zwanych teraz czynnikami). Celem analizy czynnikowej jest pogrupowanie zmiennych silnie skorelowanych i stworzenie na ich podstawie mniejszej liczby czynników, przy jak najmniejszej utracie informacji. Jak widać pokrywa się to z celem PCA.

Liczba czynników

- wykres osypiska,
- kryterium Kaisera-Guttmana – liczba czynników równa liczbie wartości własnych większych od 1,
- analiza równoległa,
- współrzędne optymalne,
- czynnik przyspieszenia.

Struktura czynników

Struktura ładunków w FA nie jest jednoznaczna, istnieje nieskończenie wiele rozwiązań dających identyczne powiązania pomiędzy oryginalnymi zmiennymi i czynnikami. Przeprowadza się zatem rotację czynników, w taki sposób, aby jedna zmienna nie wchodziła z dużym ładunkiem do więcej niż jednego czynnika. Mamy dwa typy rotacji: ortogonalną i skośną. W przypadku tej pierwszej otrzymujemy nieskorelowane czynniki, ta druga dopuszcza czynniki skorelowane. Zastosowanie rotacji ortogonalnej prowadzi do łatwiejszej interpretacji wyników (ładunki są w tym przypadku korelacjami pomiędzy czynnikami, a oryginalnymi zmiennymi), natomiast rotacje skośne prowadzą najczęściej do modelu o nieco lepszym dopasowaniu. Na rotacje można patrzeć jak na obroty osi w celu jak najprostszego w interpretacji ułożenia punktów, jeśli dopuścimy osie nieprostopadłe, to mamy rotację skośną.

Popularne rotacje ortogonalne

- varimax – otrzymujemy czynniki z kilkoma dużymi ładunkami (reszta ładunków jest bliska 0). W efekcie otrzymujemy czynniki, które są mocno skorelowane z małą liczbą zmiennych i praktycznie nieskorelowane z pozostałymi.
- quartimax – każda zmienna jest mocno skorelowana jedynie z jednym czynnikiem i wcale (lub prawie wcale) z innymi.

Popularne rotacje skośne

- oblimin – wykorzystywane jest kryterium z metody varimax, przy czym czynniki mogą być skośne (kontroluje to specjalny parametr).
- promax – wykorzystywana jest rotacja ortogonalna (najczęściej varimax), ładunki podnoszone są do pewnych potęg. Celem jest uzyskanie rozwiązania jak najbardziej dopasowanego przy wykorzystaniu najmniejszej możliwej potęgi i o minimalnie skorelowanych czynnikach. Metoda ta jako znacznie szybsza od metody oblimin i znajduje zastosowanie w przypadku dużych zbiorów danych.

W praktyce najczęściej wykorzystywana jest rotacja typu varimax.

Analiza czynnikowa i analiza składowych głównych – różnice

- W przypadku analizy czynnikowej zakłada się, że wariancja każdej zmiennej może być podzielona na wariancję wspólną (dzieloną z innymi zmiennymi) i wariancję swoistą (charakterystyczną dla danej zmiennej). Analiza czynnikowa bada jedynie wariancję wspólną, podczas gdy analiza składowych głównych całkowitą wariancję. Co za tym idzie celem PCA nie jest wyjaśnienie korelacji pomiędzy zmiennymi lecz objaśnienie wariancji danych, natomiast FA dąży do wyjaśnienia kowariancji (korelacji).

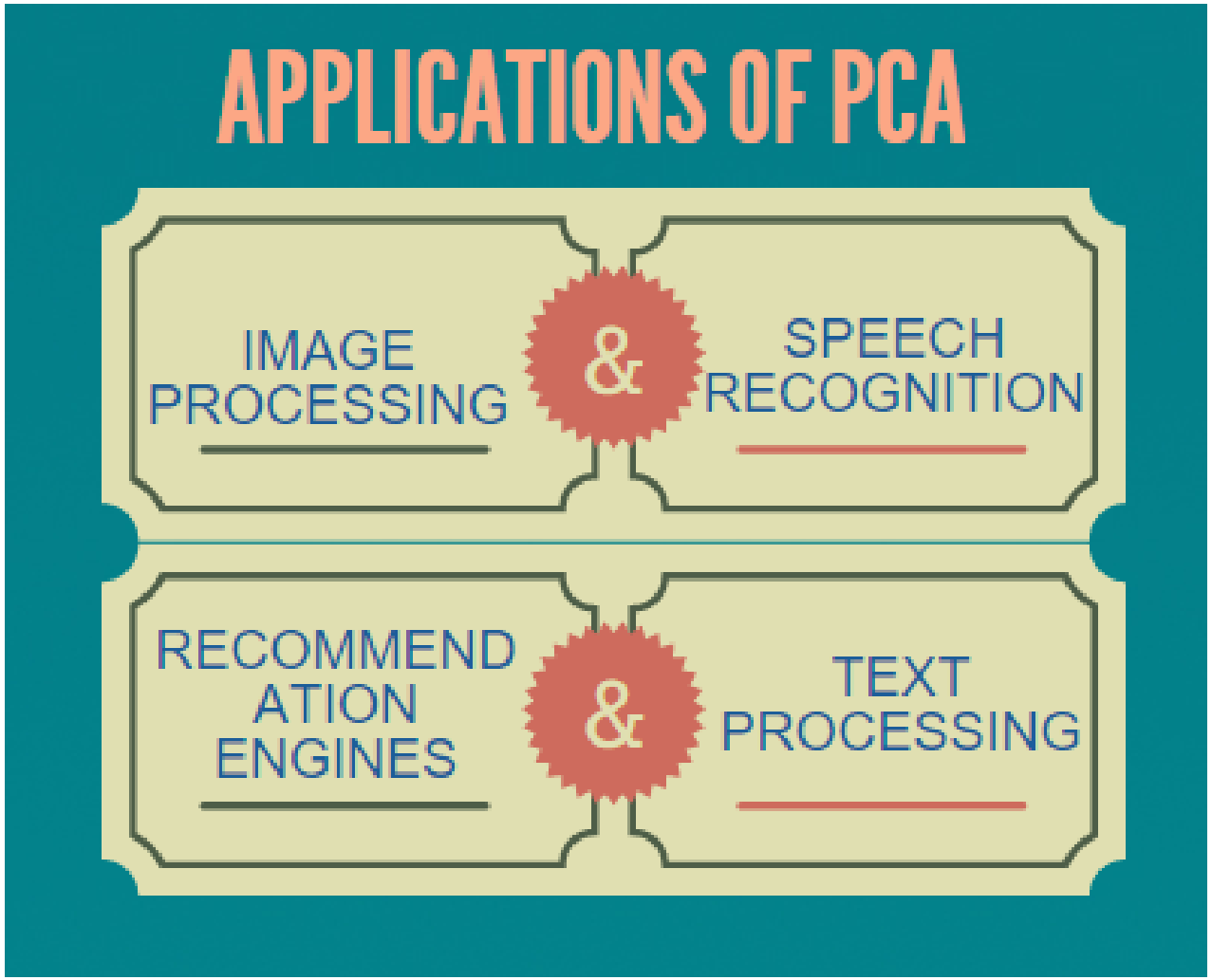
Analiza czynnikowa i analiza składowych głównych – różnice

- Składowe wyznaczone z macierzy korelacji i kowariancji istotnie różnią się w przypadku analizy składowych głównych, w przypadku analizy czynnikowej są takie same (jest niezmiennicza ze względu na skalowanie).
- W przypadku PCA wraz z dodaniem kolejnej składowej do rozwiązania, poprzednie składowe nie zmieniają się. Nie ma to miejsca w przypadku analizy czynnikowej (dodanie kolejnego czynnika zmienia pozostałe).
- Wyznaczenie składowych głównych jest znacznie prostsze od wyznaczenia czynników.

Analiza czynnikowa i analiza składowych głównych – zastosowanie

Analiza składowych głównych jest preferowana jako metoda redukcji danych, podczas gdy analiza czynnikowa jest stosowana gdy celem jest wykrycie struktury zjawiska. Podobnie jak w przypadku PCA oryginalne zmienne przedstawiane są jako kombinacje liniowe. Współczynniki tych kombinacji nazywane są jak poprzednio ładunkami i ich interpretacja jest analogiczna. Uzyskiwane wyniki są często bardzo zbliżone, zwłaszcza jeśli wariancje są małe. Jeśli zmienne są nieskorelowane to obie metody są bezużyteczne.

Analiza czynnikowa i analiza składowych głównych – zastosowanie



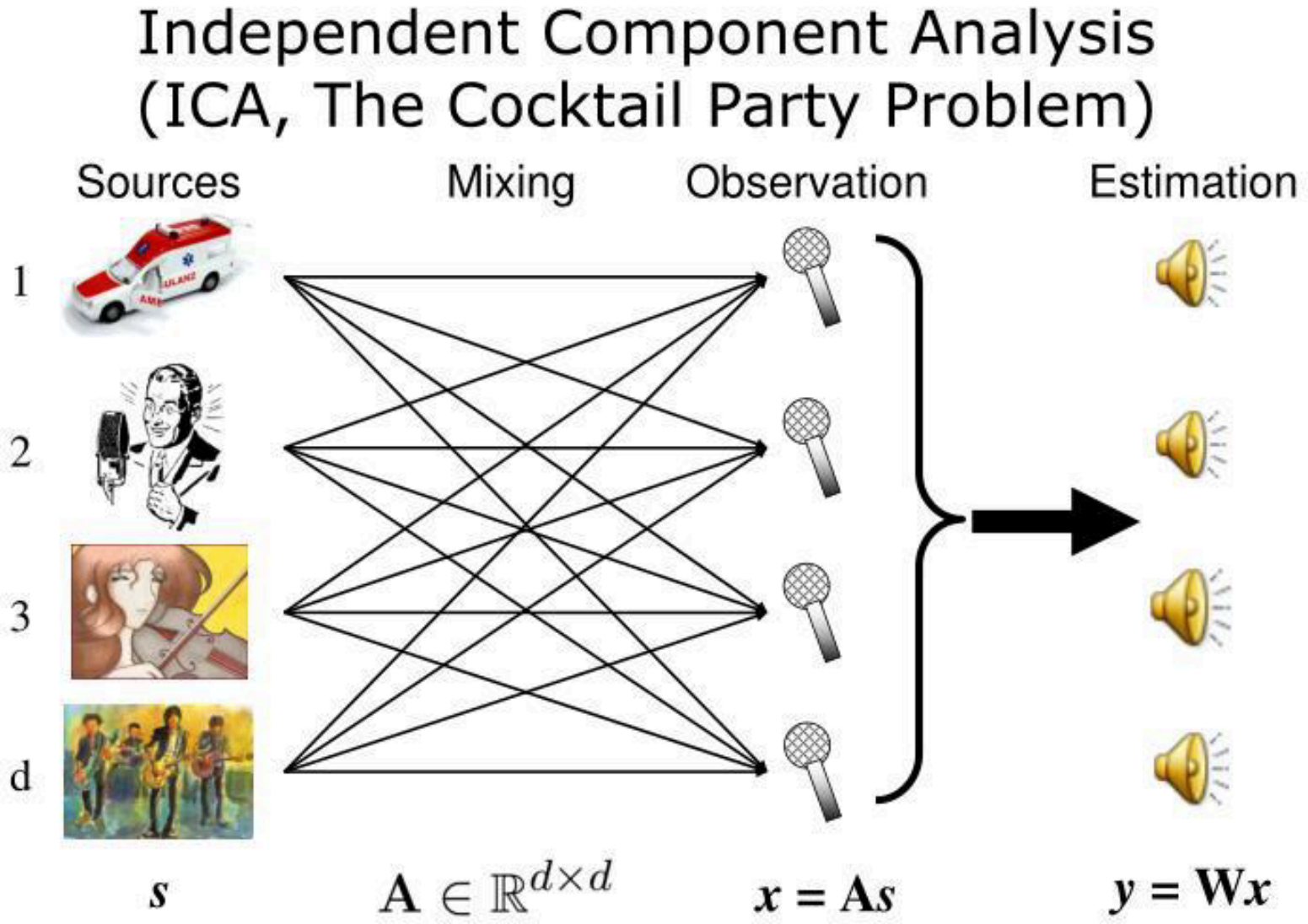
Analiza składowych niezależnych

Analiza składowych niezależnych (*ang. independent component analysis (ICA)*) jest metodą podobną do PCA, aczkolwiek ma więcej wspólnego z teorią informacji niż ze statystyką. O ile PCA konstruuje składowe nieskorelowane, to ICA stara się odnaleźć składowe niezależne (o ile pochodzą z rozkładów nie-normalnych).


Herault, J., Jutten, C. (1986). Space or time adaptive signal processing by neural models. *Proceedings AIP Conference: Neural Networks for Computing* 151: 206–211.


Tharwat, A. (2018). Independent component analysis: An introduction. *Applied Computing and Informatics*.
https://doi.org/10.1016/j.aci.2018.08.006

Analiza składowych niezależnych



t-SNE

Metoda t-SNE to stochastyczna metoda porządkowania sąsiadów w oparciu o rozkład t (*ang. t-Distributed Stochastic Neighbor Embedding*). Jest to nieliniowa i nienadzorowana technika stosowana przede wszystkim do eksploracji i wizualizacji danych wielowymiarowych.

van der Maaten, L.J.P., Hinton, G.E. (2008). Visualizing Data Using t-SNE. Journal of Machine Learning Research 9:2579–2605.

t-SNE

t-SNE vs PCA

1. t-SNE jest metodą probabilistyczną. Czasami w t-SNE różne przebiegi z tymi samymi hiperparametrami mogą dawać różne wyniki, podczas gdy dla PCA zawsze będzie to ten sam wynik.
2. PCA jest techniką liniowej redukcji wymiarów, która dąży do maksymalizacji wariancji. Zatem polega gównie na tym aby różne punkty umieszczać daleko od siebie w reprezentacji niższego wymiaru. Może to prowadzić do kiepskiej wizualizacji szczególnie w przypadku nieliniowych struktur.
3. t-SNE w przeciwieństwie do PCA zachowuje odległości pomiędzy parami odwzorowując nieliniowość i jest w stanie zinterpretować złożoną zależności pomiędzy cechami.
4. t-SNE jest drogi obliczeniowo. W przypadku większych próbek i dużej liczbie wymiarów wyliczenie t-SNE może potrwać nawet kilka godzin, podczas gdy PCA zakończy się w kilka sekund lub minut.

UMAP

Metoda UMAP (*ang. Uniform Manifold Approximation and Projection*) jest nieliniowym rozszerzeniem metody PCA. Jest jednak dużo bardziej efektywna i dokładna. UMAP produkuje podobne lub lepsze reprezentacje do t-SNE, jako że zachowuje więcej globalnych cech danych, i jest stabilniejszy. Ponadto UMAP jest wydajniejszy od t-SNE.
W pierwszym kroku wyliczane są odległości w wysoko wymiarowej (oryginalnej) przestrzeni, następnie są rzutowane na niżej wymiarową przestrzeń i wyznaczane są odległości między punktami w tej nowej przestrzeni. Następnie używana jest metoda gradientu stochastycznego aby zminimalizować różnice pomiędzy tymi odległościami.

McInnes, L., Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426v2

Idea

Jedną z wad metody PCA jest możliwość używania jedynie zmiennych ilościowych, kolejnym konieczność posiadania pełnych danych z doświadczenia (nie da się użyć PCA jeśli mamy wyłącznie informacje o podobieństwie obiektów). Skalowanie wielowymiarowe (*ang. multidimensional scaling (MDS)*) pozbawione jest tych wad. Jest to metoda redukcji wymiarowości bazująca na macierzy niepodobieństwa pomiędzy obiektami ($\boldsymbol{D} = (d_{ij})$). Celem tej metody jest wyznaczenie współrzędnych obserwacji w nowym układzie $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, w taki sposób aby odległości pomiędzy obiektami w tym nowym układzie współrzędnych były maksymalnie podobne do oryginalnych odległości pomiędzy obserwacjami:

$$d_{ij} \approx ||\boldsymbol{x}_i - \boldsymbol{x}_j||.$$

Funkcja stresu

Funkcję oceniającą rozbieżność pomiędzy danymi niepodobieństwami, a obliczonymi w nowej przestrzeni danych nazywamy funkcją stresu:

$$\text{Stress}_{\boldsymbol{D}}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = \sqrt{\sum_{i \neq j = 1, \ldots N} (d_{ij} - ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2)}.$$

Naszym celem jest oczywiście jej minimalizacja. Uzyskana wartość funkcji stresu może służyć za miarę jakości uzyskanego odwzorowania.

Stres	Jakość dopasowania
$\geq$ 0,20	Słabe
0,10-0,20	Przeciętne
0,05-0,10	Dobre
0,025-0,05	Doskonałe
0,0-0,025	Idealne

Rodzaje skalowania

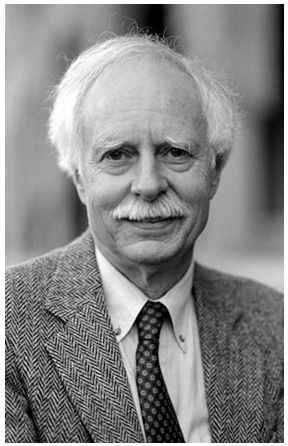
- Metryczne – minimalizujemy sumę modułów (kwadratów) różnic pomiędzy oryginalnymi odległościami oraz odległościami w nowo powstałym układzie współrzędnych. Zakładamy, że dysponujemy jedynie cechami ilościowymi. Jeśli dysponujemy oryginalnym zbiorem danych, a nie macierzą niepodobieństw, skalowanie wielowymiarowe jest tożsame z analizą składowych głównych (na macierzy kowariancji) i nazywa się klasycznym skalowaniem wielowymiarowym (*ang. principal coordinates analysis*). Skalowanie metryczne używamy gdy mamy przekonanie, że konkretna odległość w sposób właściwy reprezentuje odległości pomiędzy obiektami.

Rodzaje skalowania

- Niemetryczne – poszukujemy optymalnego porządku pomiędzy odległościami, przy czym nie ma znaczenia sama wartość odległości, jedynie ich porządek. Dane mogą być mieszaniną danych jakościowych i ilościowych. W przypadku danych jakościowych musi istnieć pomiędzy nimi pewien porządek. W przeciwieństwie do klasycznego skalowania nie istnieje analityczne rozwiązanie tego zagadnienia. Co gorsza procedura poszukiwania rozwiązania jest iteracyjna i wymaga początkowej konfiguracji punktów. Ta metoda jest częściej używana w praktyce.

Historia

Joseph Kruskal (1928-2010)



Roger Shepard (1895-1973)

Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1–27.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika* 29:115–129.

Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika* 27:125–140.

Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika* 27:219–246.

Torgerson, W.S. (1958). *Theory and methods of scaling*. Wiley.

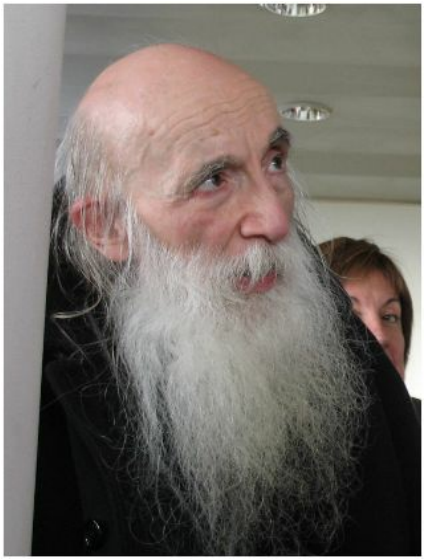
Idea

Analiza korespondencji (*ang. correspondence analysis (CA)*) to technika, która pozwala graficznie przedstawić w niskowymiarowej przestrzeni dane (najczęściej jakościowe) zawarte w tablicy wielodzielczej. Stosowana jest szczególnie często w naukach biologicznych oraz społecznych, z uwagi na często występujące macierze kontyngencji. Jeśli stwierdzimy zależność badanych cech możemy przejść do właściwej analizy korespondencji. Chcemy dokonać rzutowania oryginalnych danych na przestrzeń o jak najmniejszym wymiarze, przy czym w tej nowej przestrzeni powinna być zachowana możliwie najlepiej odległość χ^2. O tym ile oryginalnej odległości zostało zachowane mówi tzw. inercja:

$$\text{Inercja} = \frac{\chi^2}{N}.$$

Idea

Analiza korespondencji jest właściwie metodą metrycznego skalowania wielowymiarowego z odległością χ^2 jako niepodobieństwem.



Jean-Paul Benzécri (1932-2019)

Benzécri, J.P. (1973). *L'Analyse des Données. Volume II. L'Analyse des Correspondances*. Dunod.

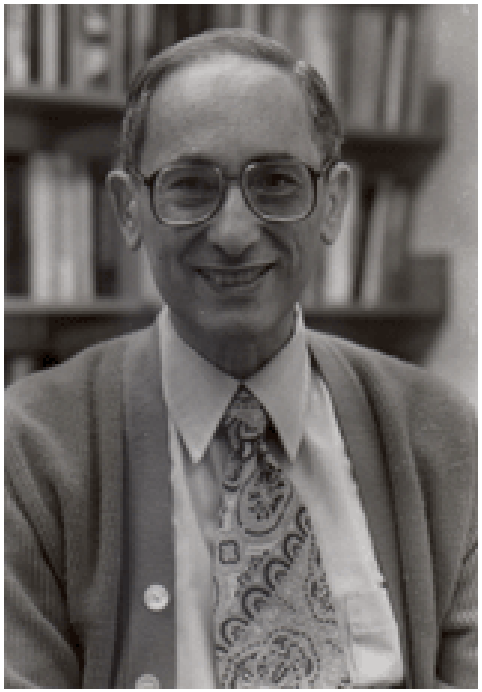
Hirschfeld, H.O. (1935). A connection between correlation and contingency. *Proc. Cambridge Philosophical Society* 31:520-524.

Wykresy obrazkowe

Czasami do wizualizacji danych wystarczają bardzo proste metody, które pomagają raczej we wstępnej analizie danych. Tego typu wykresy to wielowymiarowe wykresy obrazkowe. Podstawową ich ideą jest przedstawienie pojedynczych obserwacji za pomocą obiektów graficznych, których własności przypisano do zmiennych. Tak skonstruowane obiekty są unikalne dla każdej konfiguracji i jako takie mogą zostać rozpoznane przez badacza w sposób wizualny.

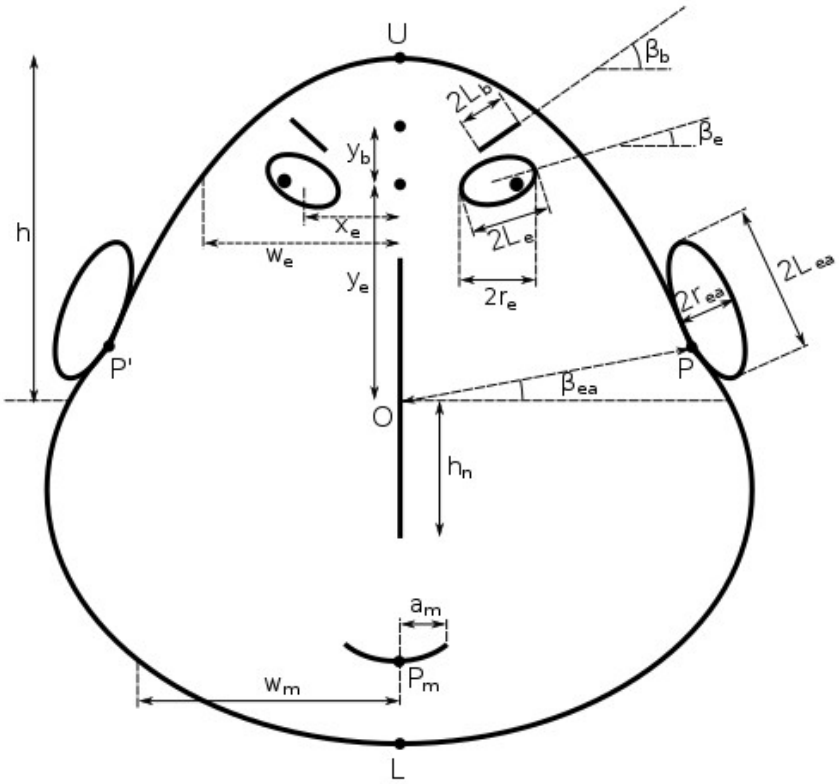
Wykresy obrazkowe

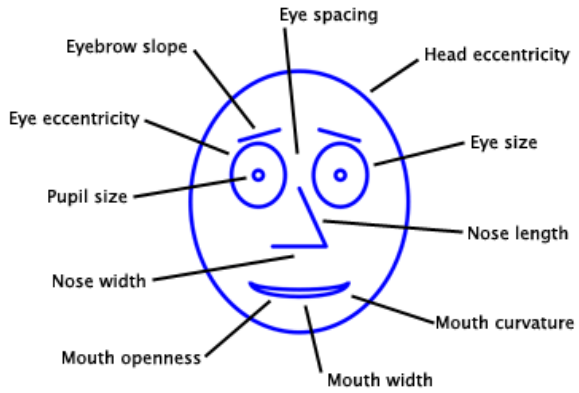
Twarze CHERNOFFA. Dla każdej obserwacji rysowany jest oddzielny obrazek „twarzy". Do kształtów i wielkości pojedynczych rysów twarzy (np. szerokość nosa, kąt brwi, wysokość uszu) przypisywane są względne wartości wybranych zmiennych. W taki sposób możemy zwizualizować do 18 cech, dodatkowe 18 uzyskamy, jeśli osobno potraktujemy lewą i prawą połowę twarzy.



Herman Chernoff (1923-)

Wykresy obrazkowe





Wykresy obrazkowe

Używane elementy twarzy: eye size (1), pupil size (2), position of pupil (3), eye slant (4), horizontal position of eye (5), vertical position of eye (6), curvature of eyebrow (7), density of eyebrow (8), horizontal position of eyebrow (9), vertical position of eyebrow (10), upper hair line (11), lower hair line (12), face line (13), darkness of hair (14), hair shading slant (15), nose (16), size of mouth (17), curvature of mouth (18).


Chernoff, H. (1973). The use of faces to represent points in *k*-dimensional space graphically *Journal of the American Statistical Association* 68(342):361-368.


Flury, B., Riedwyl, H. (1981). Graphical representation of multivariate data by means of asymmetrical faces. *Journal of the American Statistical Association* 76(376):757-765.

Wykresy obrazkowe

Wykres gwiazdowy oraz jego szczególny przypadek wykres radarowy (*ang. radar (spider) plot*). W przypadku wykresu gwiazdowego dla każdej obserwacji rysowany jest oddzielny obrazek w kształcie gwiazdy. Względne wartości wybranych zmiennych dla każdego przypadku reprezentowane są przez długości ramion gwiazdy (zgodnie z ruchem wskazówek zegara, począwszy od godziny 12:00). Końce ramion są połączone linią. W przypadku wykresu radarowego wszystkie gwiazdy nanosimy na siebie.