Rozdział 1

Metoda największej wiarygodności

Ogólnie w procesie estymacji na podstawie prób x^i (każde x^i może być wektorem) wyznaczamy parametr λ (w ogólnym przypadku również wektor) opisujący domniemany rozkład prawdopodobieństwa. Na podstawie tegoż rozkładu możemy z kolei określić *a posteriori* prawdopodobieństwo próby x^i. Logicznym wydaje się postulat, aby parametr(y) λ dobierać tak, aby zmaksymalizować prawdopodobieństwo *a posteriori* prób, z których je wyznaczamy. Funkcją wiarygodności nazywać możemy iloczyn prawdopodobieństwa *a posteriori* dla N dostępnych prób

$$L = \prod_{i=1}^{N} f(x_i, \lambda); \quad l = \ln(L) = \sum_{i=1}^{N} \ln f(x_i, \lambda) \tag{1.1}$$

Szukamy jej maksimum, czyli (zwykle) zera pochodnej.

Przykład 1

Wyznaczanie stałej fizycznej na podstawie N różnych eksperymentów (o różnej dokładności). Niech błędy podlegają rozkładowi Gaussa.

Estymowany parametr λ to wartość oczekiwana stałej. Prawdopdobieństwo *a posteriori* wyniku x_i eksperymentu o danej wariancji σ_i^2

$$f(x^i, \lambda) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-(x^i-\lambda)^2}{2\sigma_i^2}} \tag{1.2}$$

Funkcja wiarygodności

$$L = \prod_{i=1}^{N} f(x_i, \lambda) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-(x^i-\lambda)^2}{2\sigma_i^2}} \tag{1.3}$$

a jej logarytm

$$l = -\sum_{i=1}^{N} \frac{(x_i-\lambda)^2}{2\sigma_i^2} - \sum_{i=1}^{N} \ln(\sqrt{2\pi}\sigma_j) \tag{1.4}$$

Maksimum przewidujemy zwykle w zerze pochodnej

$$\frac{\delta l}{\delta \lambda} = \sum_{i=1}^{N} \frac{x_i - \lambda}{\sigma_i^2} = 0 \Rightarrow \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} = \lambda \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \Rightarrow \lambda_{NW} = \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \tag{1.5}$$

© 1999..2003 P.J. Durka

1.1 Regresja liniowa

Pary pomiarów (x_i, y_i). Dla każdego x_i, y_i traktujemy jak zmienną losową z rozkładu normalnego o wartości średniej $a + b\ x_i$ i wariancji σ_i^2. Prawdopodobieństwo *a posteriori* otrzymania N wyników y_i dla określonych x_i przy założeniu wartości a i b.

$$f(\overline{y} \mid \overline{x}, a, b) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{(y_i - a - bx_i)^2}{2\sigma_i^2}} = \frac{1}{\sqrt{(2\pi)^n}} \prod_{i=1}^{N} \frac{1}{\sigma_i} e^{-\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_i^2}} \tag{1.6}$$

logarytmiczna →**funkcja wiarygodności**:

$$l = -\frac{N}{2} \ln 2\pi + \ln(\prod_{i=1}^{N} \frac{1}{\sigma_i}) - \frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_i^2} \tag{1.7}$$

σ_i zwykle nie znamy, możemy przyjąć $\forall i \sigma_i = \sigma$. Pozostaje szukanie minimum sumy $S = \sum_{i=1}^{N} (y_i - a - bx_i)^2$, w zerze pochodnej po parametrach a i b:

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^{N} (y_i - a - bx_i), \qquad \frac{\partial S}{\partial b} = -2 \sum_{i=1}^{N} x_i (y_i - a - bx_i), \ldots \tag{1.8}$$

$$D = N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2 \tag{1.9}$$

$$a = \frac{\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{D}, \qquad b = \frac{N \sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i \right) \left(\sum_{i=1}^{N} y_i \right)}{D} \tag{1.10}$$

lub:

$$b = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N} (x_i - \overline{x})^2}, \qquad a = \overline{y} - b\overline{x} \tag{1.11}$$

© 1999..2003 P.J. Durka

1.1.1 Interpretacja współczynnika korelacji

Rozważmy wariancję zmiennej y z poprzedniego rozdziału. Niech $y_i^p = a + bx_i$

$$\sum_{i=1}^{N}(y_i-\overline{y})^2=\sum_{i=1}^{N}(y_i-y_i^p+y_i^p-\overline{y})^2=\sum_{i=1}^{N}(y_i-y_i^p)^2+\sum_{i=1}^{N}(y_i^p-\overline{y})^2+2\sum_{i=1}^{N}(y_i-y_i^p)(y_i^p-\overline{y})$$

Całkowitą wariancię zmiennej y podzieliliśmy na dwa człony: wariancję estymaty y_i^p wokół wartości średniej $\overline{y}$ i wariancję obserwowanych y_i wokół estymaty y_i^p (trzeci człon znika).

Współczynnik korelacji możemy estymować jako

$$\rho^2=\frac{\left(\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})\right)^2}{\sum_{i=1}^{N}(x_i-\overline{x})^2\sum_{i=1}^{N}(y_i-\overline{y})^2} \tag{1.12}$$

Rozważmy

$$\sum_{i=1}^{N}(y_i^p-\overline{y})^2=b^2\sum_{i=1}^{N}(x_i-\overline{x})^2=\frac{\left(\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})\right)^2}{\left(\sum_{i=1}^{N}(x_i-\overline{x})^2\right)^2}\sum_{i=1}^{N}(x_i-\overline{x})^2= \tag{1.13}$$

$$=\frac{\left(\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})\right)^2}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\frac{\sum_{i=1}^{N}(y_i-\overline{y})^2}{\sum_{i=1}^{N}(y_i-\overline{y})^2}=\rho^2\sum_{i=1}^{N}(y_i-\overline{y})^2 \tag{1.14}$$

czyli

$$\rho^2=\frac{\sum_{i=1}^{N}(y_i^p-\overline{y})^2}{\sum_{i=1}^{N}(y_i-\overline{y})^2} \tag{1.15}$$

© 1999..2003 P.J. Durka

© 1999..2003 P.J. Durka

Rozdział 2

Analiza Wariancji

2.1 Rozkład F

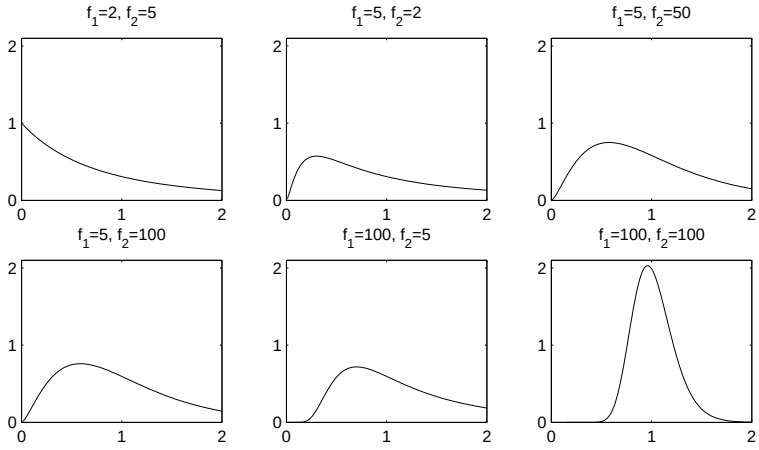
Niech zmienne x i y mają rozkłady χ^2 o odpowiednio f_1 i f_2 stopniach swobody. Zmienna

$$F = \frac{\frac{1}{f_1}x}{\frac{1}{f_2}y} = \frac{f_2 x}{f_1 y} \tag{2.1}$$

posiada rozkład F z f_1 i f_2 stopniami swobody (wartość oczekiwana $E(f) = \frac{f_2}{(f_2-2)}$)

$$f(F) = \left(\frac{f_1}{f_2}\right)^{\frac{f_1}{2}} \frac{\Gamma\left(\frac{1}{2}(f_1+f_2)\right)}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)} F^{\frac{f_2}{2}-1}\left(1+\frac{f_1}{f_2}F\right)^{-\frac{f_1+f_2}{2}} \tag{2.2}$$

Dla próby z rozkładu normalnego wielkość



Rysunek 2.1: Rozkład F Fischera dla przykładowych ilości stopni swobody f_1 i f_2.

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{\sigma^2} \tag{2.3}$$

podlega rozkładowi χ^2 o $f = N-1$ stopniach swobody. Jeśli dwie takie próby zostały pobrane z jednej populacji, to iloraz

$$F = \frac{(N_y - 1)\sum_{i=1}^{N}(x_i - \overline{x})^2}{(N_x - 1)\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{2.4}$$

podlega rozkładowi F o f_y i f_x stopniach swobody.

© 1999..2003 P.J. Durka

2.2 Analiza wariancji *(ANalysis of VAriance - ANOVA)*

N obserwacji $\{x_i\}_{i=1..N}$ podzielonych na k grup wedle jakiegoś kryterium: $N = n_1+n_2+...+n_k$. Średnie wewnątrz grup

$$\overline{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \tag{2.5}$$

Rozważmy sumę kwadratów odchyleń wszystkich elementów próby od wartości średniej całej próby:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i+\overline{x}_i-\overline{x})^2 = \tag{2.6}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x}_i-\overline{x})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)(\overline{x}_i-\overline{x})$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)(\overline{x}_i-\overline{x}) = \sum_{i=1}^{k}(\overline{x}_i-\overline{x})\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i) = 0 \tag{2.7}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x}_i-\overline{x})^2 = \sum_{i=1}^{k} n_i(\overline{x}_i-\overline{x}) \tag{2.8}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 + \sum_{i=1}^{k} n_i(\overline{x}_i-\overline{x})^2 = s^2_{wew} + s^2_{pom} \tag{2.9}$$

Jeśli wszystkie pomiary pochodzą z tej samej populacji o wariancji σ^2, to

$$\frac{s_{wew}}{\sigma^2} \;\; i \;\; \frac{s_{pom}}{\sigma^2} \tag{2.10}$$

podlegają rozkładom χ^2 o odpowiednio $n-k$ i $k-1$ stopniach swobody. Iloraz

$$\frac{(n-k)\,s_{pom}}{(k-1)\,s_{wew}} \tag{2.11}$$

podlega rozkładowi F o $k-1$ i $n-k$ stopniach swobody. Wyrażenia

$$\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 \;\; oraz \;\; \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x}_i-\overline{x})^2 \tag{2.12}$$

czyli

$$\frac{s^2_{wew}}{n-k} \;\; oraz \;\; \frac{s^2_{pom}}{k-1} \tag{2.13}$$

są nieobciążonymi estymatami wariancji populacji.

© 1999..2003 P.J. Durka

Rozdział 3

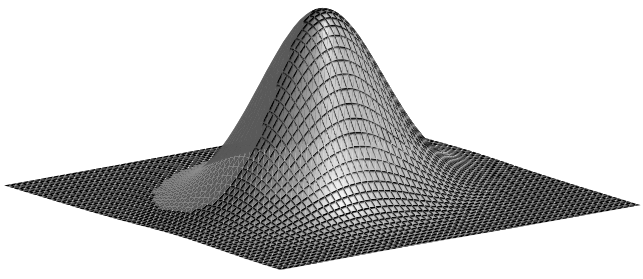
Elementy statystyki wielowymiarowej

3.1 Dwumianowy rozkład normalny

$$f(t) = ke^{-\frac{1}{2}(t-\mu)A(t-\mu)^T} \tag{3.1}$$

$t = (x, y)-$ wektor zmiennej losowej
$\mu = (\mu_1, \mu_2)-$ wektor wartości oczekiwanych
$k-$ stała normalizjąca
$A-$ odwrotność macierzy kowariancji C

$$A = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}^{-1} = \frac{1}{\sigma_x^2\sigma_y^2 - (\sigma_{xy})^2} \begin{bmatrix} \sigma_y^2 & -\sigma_{xy} \\ -\sigma_{xy} & \sigma_x^2 \end{bmatrix} \tag{3.2}$$



Rysunek 3.1: Dwumianowy rozkład normalny, wartość prawdopodobieństwa jako wysokość nad płaszczyzną

3.2 Macierz kowariancji

$$C = E[(x-\mu)(x-\mu)^T], \qquad c_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] \tag{3.3}$$

© 1999..2003 P.J. Durka

dla $x = (x_1, x_2)$ i $\mu = (\mu_1, \mu_2)$

$$C = E\left[\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}(x_1 - \mu_1, x_2 - \mu_2)\right] = \begin{bmatrix} E\left[(x_1 - \mu_1)^2\right] & E\left[(x_1 - \mu_1)(x_2 - \mu_2)\right] \\ E\left[(x_2 - \mu_2)(x_1 - \mu_1)\right] & E\left[(x_2 - \mu_2)^2\right] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \quad (3.4)$$

Interpretacja współczynnika korelacji znajduje się pod hasłem *Regresja liniowa*.

3.3 Analiza wariancji wielu zmiennych (*Multivariate ANalysis of VAriance - MANOVA*)

Zmienna losowa X opisywana wektorem $(x_1, ..., x_k)$, podobnie wartość średnia staje się wektorem o tym samym wymiarze: $(\mu_1, ..., \mu_k)$. Macierz kowariancji (zdefiniowana w 3.2) zmiennej losowej

$$S = \begin{bmatrix} var(x_1) & cov(x_1, x_2) & ... & cov(x_1, x_k) \\ cov(x_2, x_1) & var(x_2) & ... & cov(x_2, x_k) \\ ... & ... & ... & ... \\ cov\,(x_k, x_1) & cov(x_k, x_2) & ... & var(x_k) \end{bmatrix} \quad (3.5)$$

gdzie:

$$var(x_i) = \sigma_{x_i} = E((x_i - \mu_i)^2) \quad (3.6)$$

$$cov(x_i, x_k) = \sigma_{x_i, x_k} = E((x_i - \mu_i)(x_k - \mu_k)) \quad (3.7)$$

Wielowymiarowy rozkład normalny

$$\Phi(X) = \frac{1}{\sqrt{(2\pi)^k}\sqrt{|S|}} e^{-\frac{(X-\mu)' S^{-1} (X-\mu)}{2}} \quad (3.8)$$

Od tego momentu przyjmujemy, że zmienne X podlegają takiemu właśnie rozkładowi. Jeśli X pochodzą z próby podzielonej na grupy, to podobnie jak w $ANOVA$ możemy skonstruować macierze wariancji wewnątrzgrupowych i międzygrupowych i dowieść, że $S = S_{wew} + S_{pom}$. Iloraz wyznaczników macierzy S_{wew} i S podlega rozkładowi Λ Wilksa: $\Lambda = \frac{|S_{wew}|}{|S|} = \frac{|S_{wew}|}{|S_{wew} + S_{pom}|}$ i służy testowaniu hipotezy o braku różnic między grupami. Ogólnie statystyki Wilksa można używać do testowania hipotezy h w postaci $\Lambda = \frac{|S_{wew}|}{|S_{wew} + S_h|}$, gdzie S_h — macierz kowariancji odpowiadająca testowanej hipotezie.

© 1999..2003 P.J. Durka

Rozdział 4

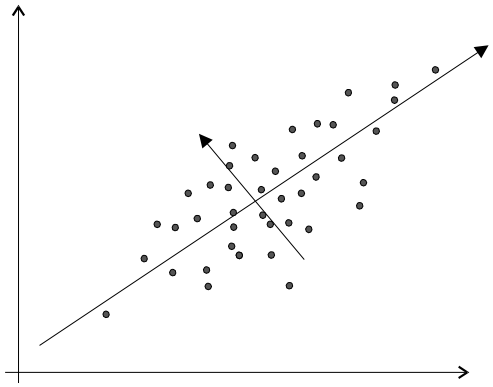
Niektóre z pominiętych haseł — w skrócie:

4.1 Analiza składowych głównych (*Principal Components Analysis, PCA*)

Przedstawiamy macierz kowariancji w postaci diagonalnej

$$S = \begin{bmatrix} r_{11} & r_{12} & ... & r_{1k} \\ r_{21} & r_{22} & ... & r_{2k} \\ ... & ... & ... & ... \\ r_{k1} & r_{k2} & ... & r_{kk} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_k \end{bmatrix} \begin{bmatrix} r_{11} & r_{21} & ... & r_{k1} \\ r_{12} & r_{22} & ... & r_{k2} \\ ... & ... & ... & ... \\ r_{1k} & r_{2k} & ... & r_{kk} \end{bmatrix} \tag{4.1}$$

Wielkości λ_i są rozwiązaniami równania $|S - \lambda I| = 0$, a wektor r_i osiami nowego układu współrzędnych. Składowe PCA są liniowymi kombinacjami obserwowanych zmiennych.



Rysunek 4.1: Kierunki składowych głównych (PCA) w dwóch wymiarach

4.1.1 Analiza dyskryminacyjna (*Discriminant Analysis*)

Wielowymiarowe wektory próby X mamy podzielone na grupy, szukamy funkcji najlepiej je rozdzielającej, która umożliwi zaklasyfikowanie nowej obserwacji. Rozdzielenie grup odpowiada w przypadku jednowymiarowym maksymalizacji stosunku wariancji międzygrupowej do wariancji wewnątrzgrupowej

$$F = \frac{(n-k)\, s_{pom}}{(k-1)\, s_{wew}} \tag{4.2}$$

W przypadku wielowymiarowym mamy do czynienia z macierzami kowariancji; możemy rozpatrywać wielkość

$$F_a = \frac{a' S_{pom} a}{a' S_{wew} a} \tag{4.3}$$

© 1999..2003 P.J. Durka

Maksymalizacja tej wielkości względem a daje wektor własny macierzy $S_{wew}^{-1}S_{pom}$ odpowiadający największej wartości własnej. Wektory własne odpowiadające kolejnym wartościom własnym zwiemy współrzędnymi dyskryminacyjnymi, tworzącymi przestrzeń dyskryminacyjną.

4.2 Analiza czynnikowa (*Factor Analysis*)

opiera się na założeniu istnienia ukrytych czynników, stara się przedstawić obserwowane zmienne w postaci:

obserwowana zmienna = liniowa kombinacja czynników + błąd

w odróżnieniu od PCA, realizującej model

składowa = liniowa kombinacja obserwowanych zmiennych

4.3 Analiza skupień — *Cluster Analysis*

Wejściem dla tej procedury jest zestaw danych, a wyjściem ich podział na grupy. Można go zrealizować na wiele sposobów: N punktów $x^1...x^N$, z których każdy opisany jest przez k cech $x_1...x_k$.

4.3.1 Metody polegające na kolejnym łączeniu punktów

Startujemy z N klastrów jednopunktowych, w każdym kroku łączymy najbliższe. Wynikiem działania jest drzewo łączenia, na którym sami musimy wybrać ilość klastrów. Wynik zależy silnie od przyjętych definicji odległości między klastrami oraz definicji odległości między punktami.

Odległości między punktami:

- Odległość Euklidesowa $d(x,y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ (czuła na różne skale cech).
- Odległość korelacyjna $d(x,y) = 1 - \rho(x,y)$, gdzie $\rho(x,y) = \frac{\sum_{i=1}^{k}(x-\overline{x})(y-\overline{y})}{\sigma_x \sigma_y}$ (znormalizowana do przedziału (0,2), mniejsza im lepiej skorelowane punkty).

Odległości między klastrami:

- Najbliższego sąsiada *(single linkage)* - odległość między dwoma najbliższymi elementami klastrów A i B: $d(A,B) = \min_{x,y} d(x,y), \qquad x \in A,\ y \in B$
- (*complete linkage*) - odległość między dwoma najbliższymi elementami klastrów A i B: $d(A,B) = \max_{x,y} d(x,y), \qquad x \in A,\ y \in B$
- *(centroid)* - odległość między środkami klastrów,
- *(avarage)* - średnia odległości, itd...

4.3.2 Metoda K-średnich (*K - means*)

Wybieramy ilość klastrów, podział dokonywany jest w iteracyjnej procedurze dążącej do minimalizacji stosunku wariancji pomiędzy klastrami do wariancji wewnątrz klastrów - niejako $ANOVA$ bez ustalonego wstępnie przyporządkowania, maksimum F poszukiwane drogą przemieszczania elementów między klastrami.

4.4 Scyzoryk (*jacknife*)

Jest to metoda analogiczna do bootstrapu, w ktorej powtórzenia generowane są przez usuwanie z próby kolejno po jednym elemencie. W porównaniu z bootstrapem daje w niektórych przypadkach oszczędność ilości obliczeń, ale ze względu na generowanie prób o mniejszej liczebności niż oryginalna wymaga stosowania mniej intuicyjnych wzorów. Jeśli chodzi o efektywność, to w przypadku liniowym jest dokładnym przybliżeniem bootstrapu, jednak w pozostałych przypadkach może prowadzić do większych błędów.

© 1999..2003 P.J. Durka