Statystyczna analiza danych

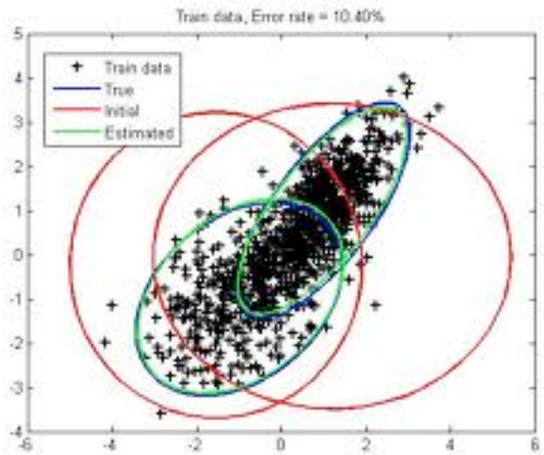
mieszanki rokładów normalnych

algorytm EM

Anna Gambin

Instytut Informatyki

Uniwersytet Warszawski



1 Wprowadzenie

2 Wielowymiarowe rozkłady normalne

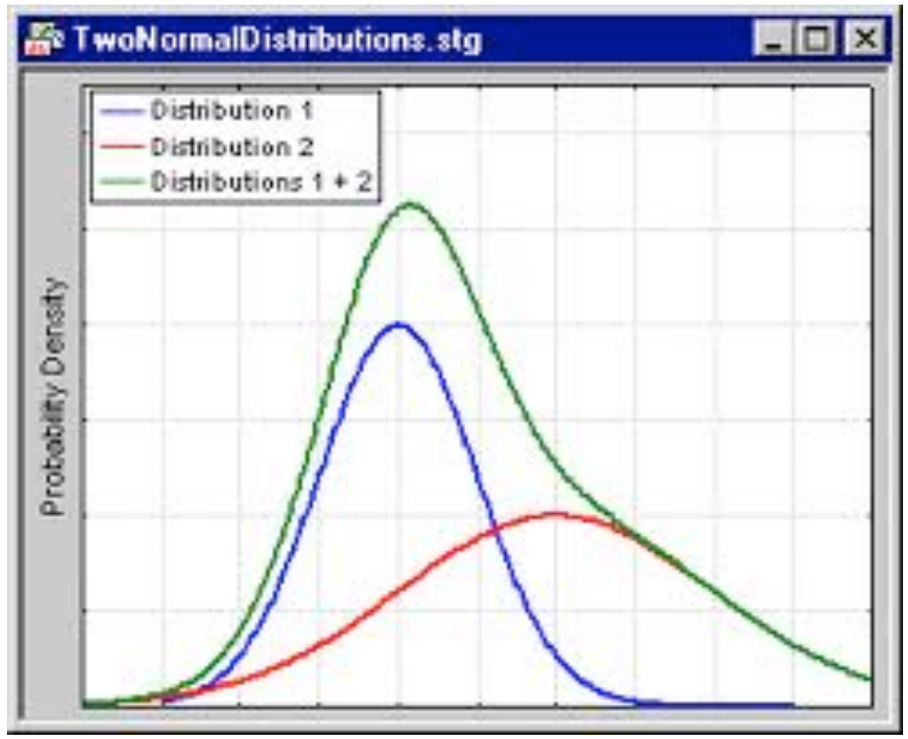
3 Algorytm k-średnich

4 Algorytm k-median

1 Wprowadzenie

Do tej pory analizowaliśmy algorytmy grupowania oparte na pojęciu podobieństwa obiektów (miarą podobieństwa była np. korelacja profili ekspresji dla genów).

Innym podejściem jest grupowanie oparte na statystycznym modelu populacji danych. Zakładamy tutaj, że obserwowane obiekty każdej grupy są niezależną próbą losową z populacji o pewnym rozkładzie, oraz że przynależność obiektów do grup została „zapomniana”.



Zaczniemy od przypomnienia podstawowych faktów dotyczących rozkładu normalnego, wzoru Bayesa i warunkowej wartości oczekiwanej.

Rozkład normalny. Rokład normalny o wartości oczekiwanej μ oraz wariancji σ^2 ma gęstość:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

Po raz pierwszy rozkład normalny był rozważany przez francuskiego matematyka Abrahama de Moivre'a w roku 1733 jako przybliżenie rozkładu dwumianowego dla dużych n. Wynik tez został później wzmocniony przez Laplace'a i jest znany jako Centralne Twierdzenie Granicze[1], które dla kompletności przytoczymy.

Twierdzenie 1 (De Moivre - Laplace) *Niech S_n oznacza liczbę sukcesów w n próbach Bernoulliego z prawdopodobieństwem sukcesu p. Wtedy dla dowolnych $0 \leq a < b \leq 1$, zachodzi:*

$$P\{a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\} \to \Phi(b) - \Phi(a) \quad dla \quad n \to \infty$$

$$gdzie \qquad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

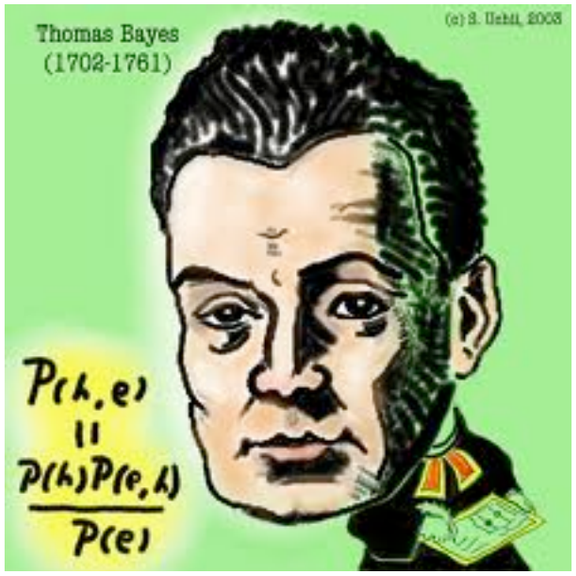
[1] uważane wraz z Silnym Prawem Wielkich Liczb za najważniejsze twierdzenie w rachunku prawdopodobieństwa.

Warunkową wartością oczekiwaną zmiennej losowej Y dla ustalonej wartości innej zmiennej $X = x$ nazywamy wartość oczekiwaną **względem rozkładu warunkowego:**

$$E(Y|X = x) = \begin{cases} \sum_k y f(y|x), \text{ jeśli Y jest zmienną dyskretną} \\ \int f(y|x)dy, \text{ jeśli Y jest zmienną ciągłą} \end{cases}$$

gdzie $f(y|x)$ jest warunkową gestością, czyli:

$$f(y|x) = \frac{f(x,y)}{f_X(x)}$$



Wzór Bayesa. Niech zdarzenia $A_1, A_2, \ldots$ spełniają następujące warunki: $A_i \cap A_j = \emptyset$ dla $i \neq j$, $A_1 \cup A_2 \cup \ldots = \Omega$ oraz $P(A_i) > 0 \quad \forall i$. Wtedy dla dowolnego zdarzenia B, jeśli $P(B) > 0$, zachodzi:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_k P(B|A_k)P(A_k)}$$

2 Wielowymiarowe rozkłady normalne

Zakładamy również, że wiemy z ilu grup pochodzą obiekty. Niech π_k będzie prawdopodobieństwem, że losowo wybrany obiekt pochodzi z populacji k o wielowymiarowym rozkładzie normalnym ze średnią $\boldsymbol{\mu}_k$ oraz macierzą kowariancji $\boldsymbol{\Sigma}_k$.

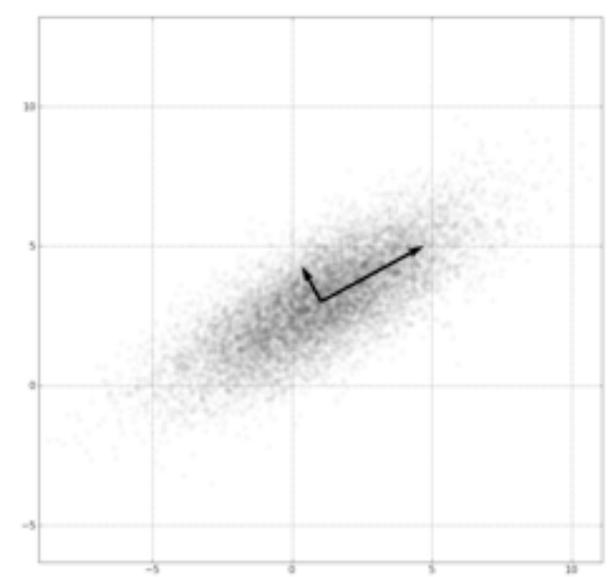
Przypomnijmy, że p-wymiarowa zmienna losowa

$$\boldsymbol{x} = (x_1, \ldots, x_p)$$

pochodząca z wielowymiarowego rozkładu normalnego ze średnią:

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$$

Multivariate (bivariate) Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the $(0.878, 0.478)$ direction and of 1 in the orthogonal direction.



i macierzą kowariancji:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_p^2 \end{bmatrix}$$

ma gęstość:

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p \cdot |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

kowariancja nie tylko dla zmiennych normalnych:

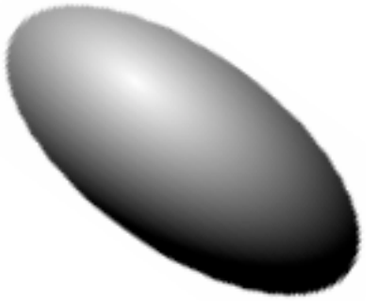
$$\mathrm{Cov}\,(x, y) = \langle (x - \overline{x})(y - \overline{y}) \rangle$$

dla wielu zmiennych losowych: macierz

$$\mathbf{C} = C_{ij} = \mathrm{Cov}\,(x_i, x_j) = \langle (x_i - \overline{x_i})(x_j - \overline{x_j}) \rangle$$

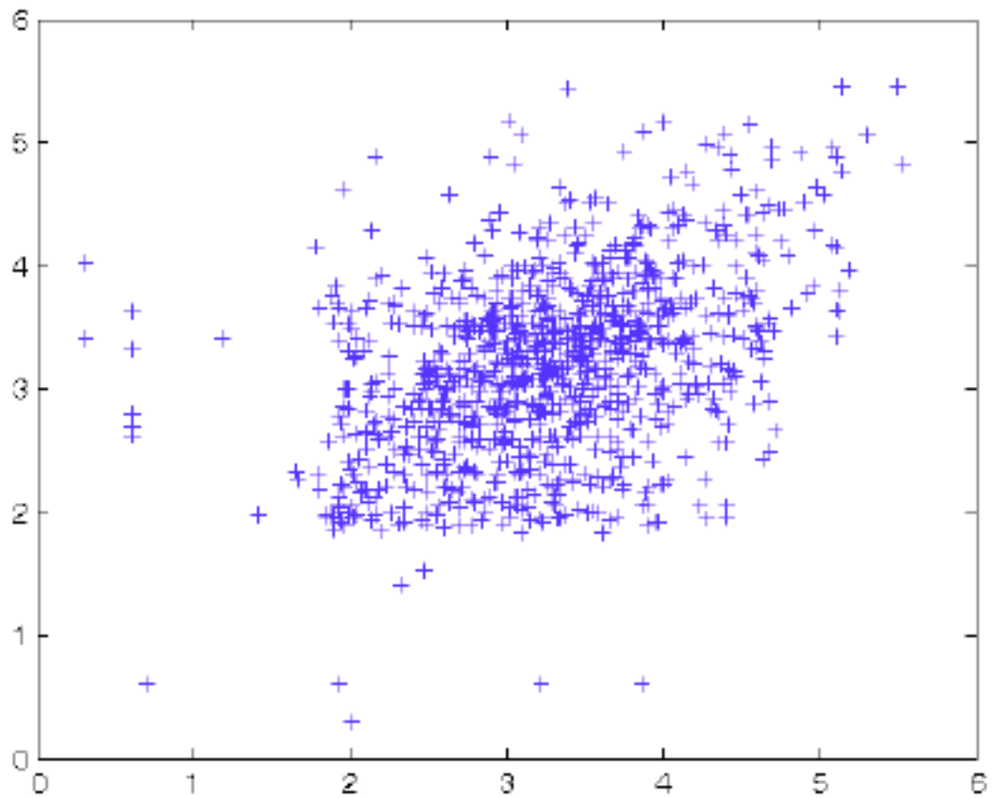
wariancja kombinacji liniowej -> forma kwadratowa:

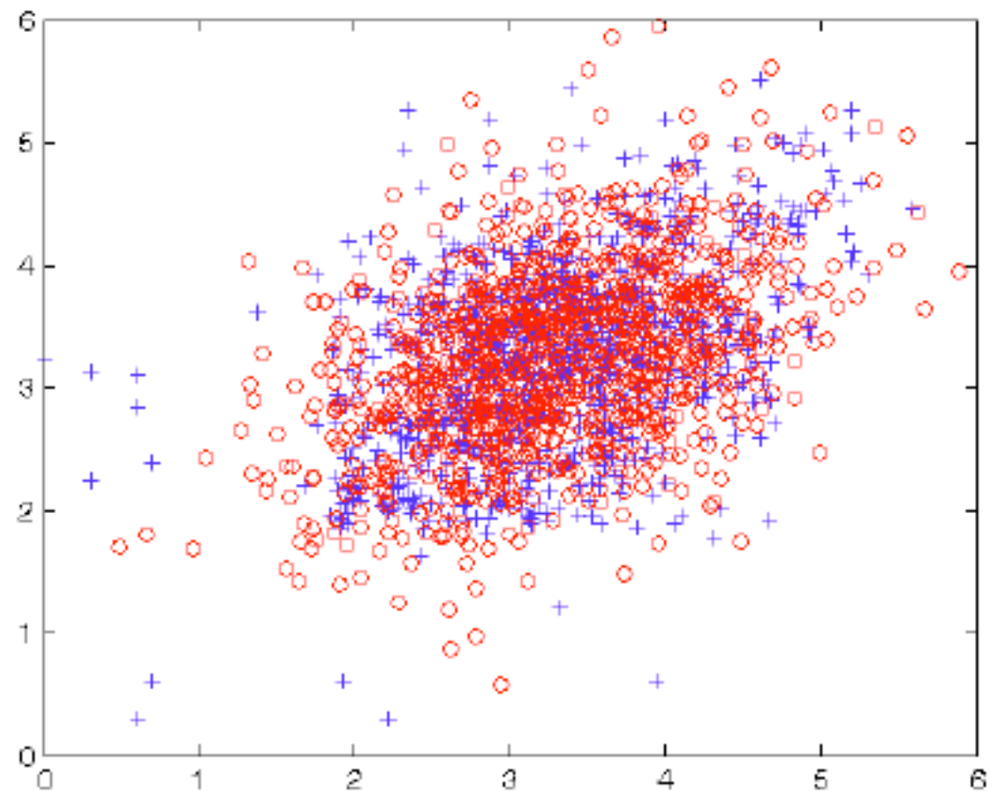
$$\begin{aligned}
\mathrm{Var}\,\left(\sum \alpha_i x_i\right) &= \left\langle \sum_i \alpha_i (x_i - \overline{x_i}) \sum_j \alpha_j (x_j - \overline{x_j}) \right\rangle \\
&= \sum_{ij} \alpha_i \left\langle (x_i - \overline{x_i})(x_j - \overline{x_j}) \right\rangle \alpha_j \\
&= \boldsymbol{\alpha}^T \mathbf{C}\, \boldsymbol{\alpha}
\end{aligned}$$

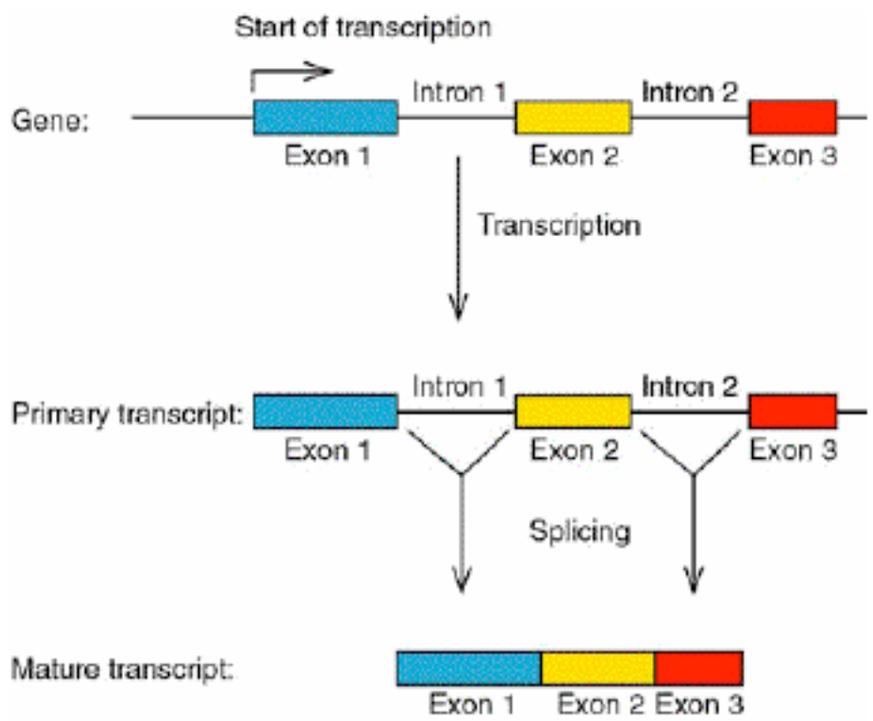


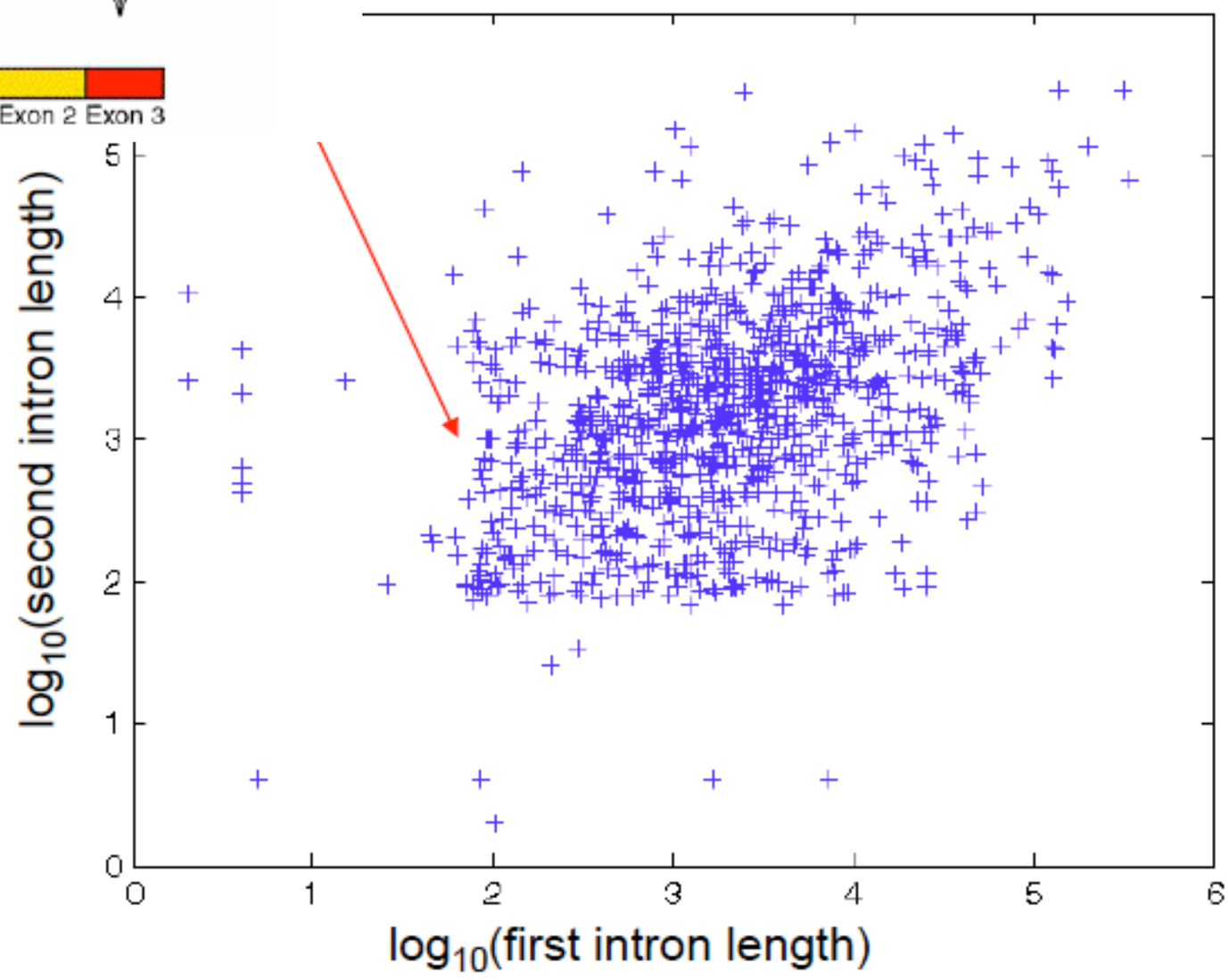
przykład biologiczny:

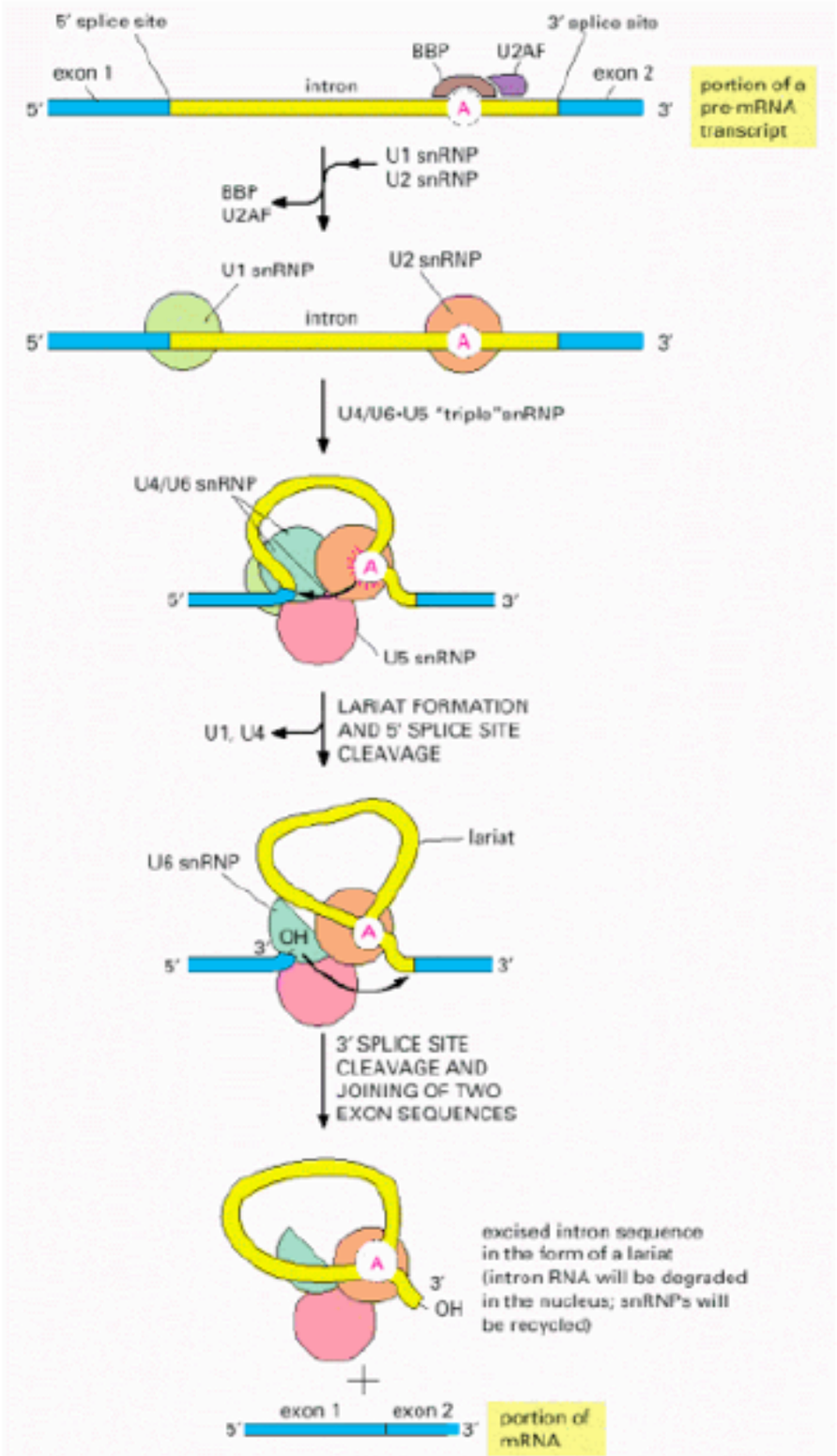
wielowymiarowy rozkład - czy normalny ???


0
1
2
3
4
5
6
0
1
2
3
4
5
6


0
1
2
3
4
5
6
0
1
2
3
4
5
6


Start of transcription
Gene:
Intron 1
Intron 2
Exon 1
Exon 2
Exon 3
Transcription
Primary transcript:
Intron 1
Intron 2
Exon 1
Exon 2
Exon 3
Splicing
Mature transcript:
Exon 1
Exon 2
Exon 3


log10(second intron length)
log10(first intron length)
0
1
2
3
4
5
6



credit: Alberts et al.
Molecular Biology of the Cell

Rozkład brzegowy dla x_j jest rozkładem normalnym ze średnią μ_j oraz wariancją σ_j^2.

Powracając do zadania grupowania, zakładamy że gęstość obiektów w grupie k wynosi:

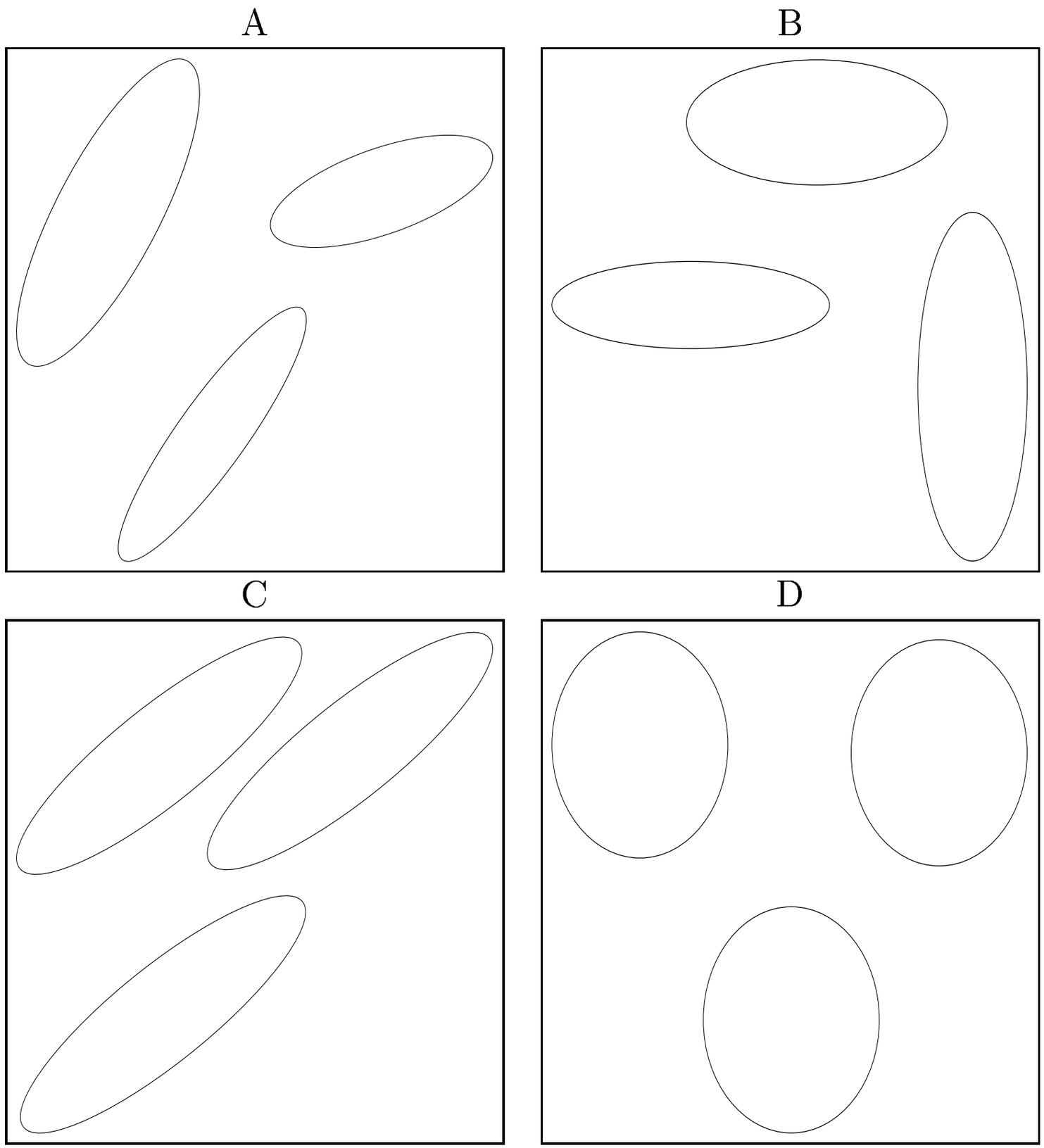
$$f_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p \cdot |\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}$$

$$\boldsymbol{\mu} = \langle \mathbf{x} \rangle \qquad \boldsymbol{\Sigma} = \left\langle (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \right\rangle$$

Często przyjmuje się dodatkowe założenie, że wszystkie grupy mają tę samą orientację oraz rozrzut, czyli $\forall k \;\; \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$. Lub jeszcze bardziej restrykcyjnie, że wszystkie grupy są sferyczne, czyli $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$.

Innym popularnym założeniem jest rozważanie diagonalnych macierzy kowariancji, czyli

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \sigma_1^{(k)^2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_p^{(k)^2} \end{bmatrix}$$



Rysunek 1: A: różne macierze kowariancji; B: diagonalne macierze kowariancji; C: identyczne macierze kowariancji; D: identyczne, sferyczne macierze kowariancji

Gęstość dla wszystkich obiektów jest *mieszaniną* gęstości dla poszczególnych grup, czyli:

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\boldsymbol{x})$$

Naszym celem jest estymacja nieznanych parametrów rozkładów. Czyni się to szeroko stosowanym algorytmem EM (*ang. expectation maximisation*) [1].

[1] A. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal Royal Stat. Soc. Series B* **39**,1-38, 1977.

Załóżmy na chwilę, że znamy „zapomniane” etykiety grup S_i $(i = 1, \ldots k)$. Wtedy *wiarygodność* naszego modelu, czyli prawdopodobieństwo, że obserwacje $x_1, \ldots x_n$ pochodzą z odpowiednich populacji wynosi:

$$\prod_{i=1}^{n} \pi_{S_i} f_{\boldsymbol{\mu}_{S_i}, \boldsymbol{\Sigma}_{S_i}}(\boldsymbol{x}_i) \tag{1}$$

Oznaczmy przez θ wektor wszystkich parametrów w naszym zadaniu:

$$\theta = (\pi_1, \ldots \pi_K, \boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots \boldsymbol{\Sigma}_K)$$

$$\prod_{i=1}^{n} \pi_{S_i} f_{\boldsymbol{\mu}_{S_i}, \boldsymbol{\Sigma}_{S_i}}(\boldsymbol{x}_i)$$

Logarytmując wiarygodność zdefiniowaną równaniem (1) otrzymujemy:

$$L(\theta) = \sum_{i=1}^{n} \log \pi_{S_i} + \log f_{\boldsymbol{\mu}_{S_i}, \boldsymbol{\Sigma}_{S_i}}(\boldsymbol{x}_i)$$

Niestety w rzeczywistości nie znamy etykiet S_i, zaś algorytm EM próbuje zmaksymalizować warunkową wartość oczekiwaną $L(\theta)$. W kroku 'E' musimy policzyć warunkową wartość oczekiwaną dla $L(\theta)$ pod warunkiem obserwowanych danych używając parametrów $\hat{\theta}^{(h-1)}$ z poprzedniego kroku. Ze wzoru Bayesa liczymy rozkład warunkowy dla S_i:

$$w_{ik}^{(h)} := \mathcal{P}_{\hat{\theta}^{(h-1)}}(S_i = k) = \frac{\hat{\pi}_k^{(h-1)} f_{\boldsymbol{\mu}_k^{(h-1)}, \boldsymbol{\Sigma}_k^{(h-1)}}(\boldsymbol{x}_i)}{\sum_{\xi=1}^{K} \hat{\pi}_\xi^{(h-1)} f_{\boldsymbol{\mu}_\xi^{(h-1)}, \boldsymbol{\Sigma}_\xi^{(h-1)}}(\boldsymbol{x}_i)} \tag{2}$$

Zatem warunkowa wartość oczekiwana wynosi:

$$\mathcal{E}_{\hat{\theta}^{(h-1)}}(L(\theta)|x_1,\ldots,x_n)) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(h)} \log \pi_k + w_{ik}^{(h)} \log f_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}) \qquad (3)$$

W kroku 'M' naszym zadaniem jest policzenie $\hat{\theta}^{(h)}$, które maksymalizuje warunkową wartość oczekiwaną 3.

Reasumując, algorytm EM powtarza kroki 'E' i 'M' dla kolejnych $h = 1, 2, \ldots$:

E: Estymuj warunkowy rozkład S_i pod warunkiem obserwowanych danych oraz parametrów rozkładów z poprzedniego kroku $\hat{\boldsymbol{\mu}}_k^{(h-1)}$ oraz $\hat{\boldsymbol{\Sigma}}_k^{(h-1)}$ dla $k = 1, 2, \ldots K$. Uzyskując wagi $w_{ik}^{(h)}$ ze wzoru (2).

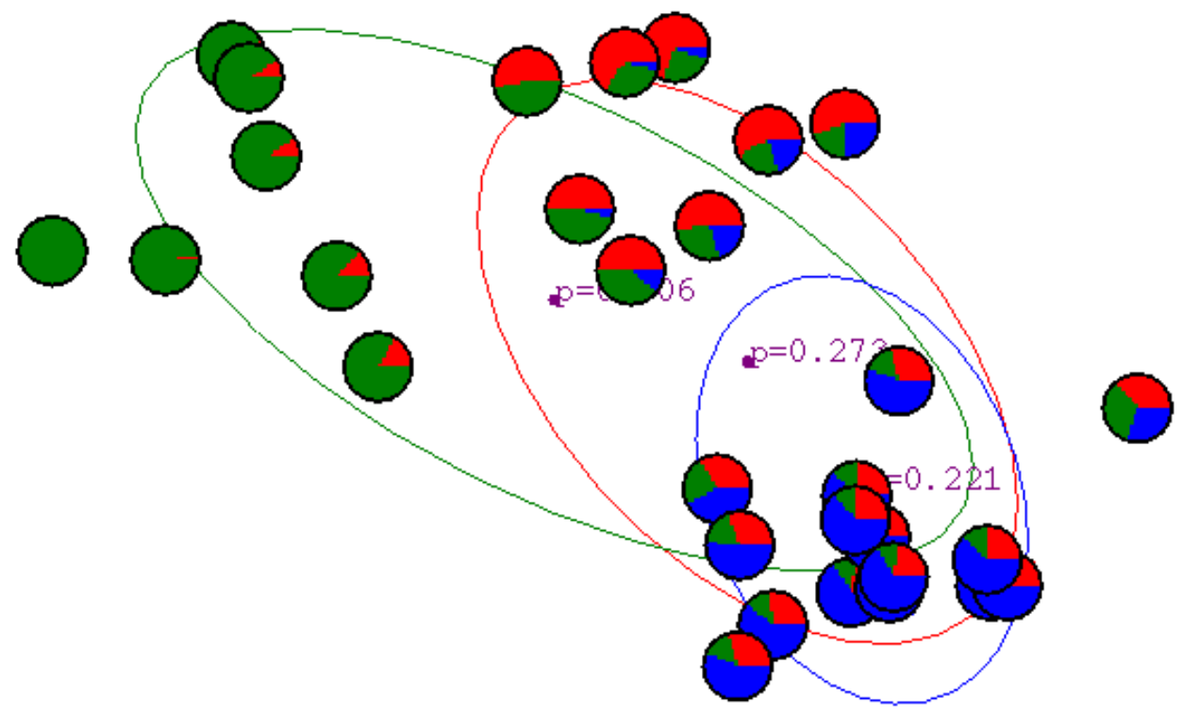
M: Popraw oczacowania parametrów następująco:

$$\hat{\pi}_k^{(h)} := \frac{1}{n} \sum_{i=1}^{n} w_{ik}^{(h)},$$

$$\hat{\boldsymbol{\mu}}_k^{(h)} := \frac{1}{\sum_{i=1}^{n} w_{ik}^{(h)}} \sum_{i=1}^{n} w_{ik}^{(h)} \boldsymbol{x}_i$$

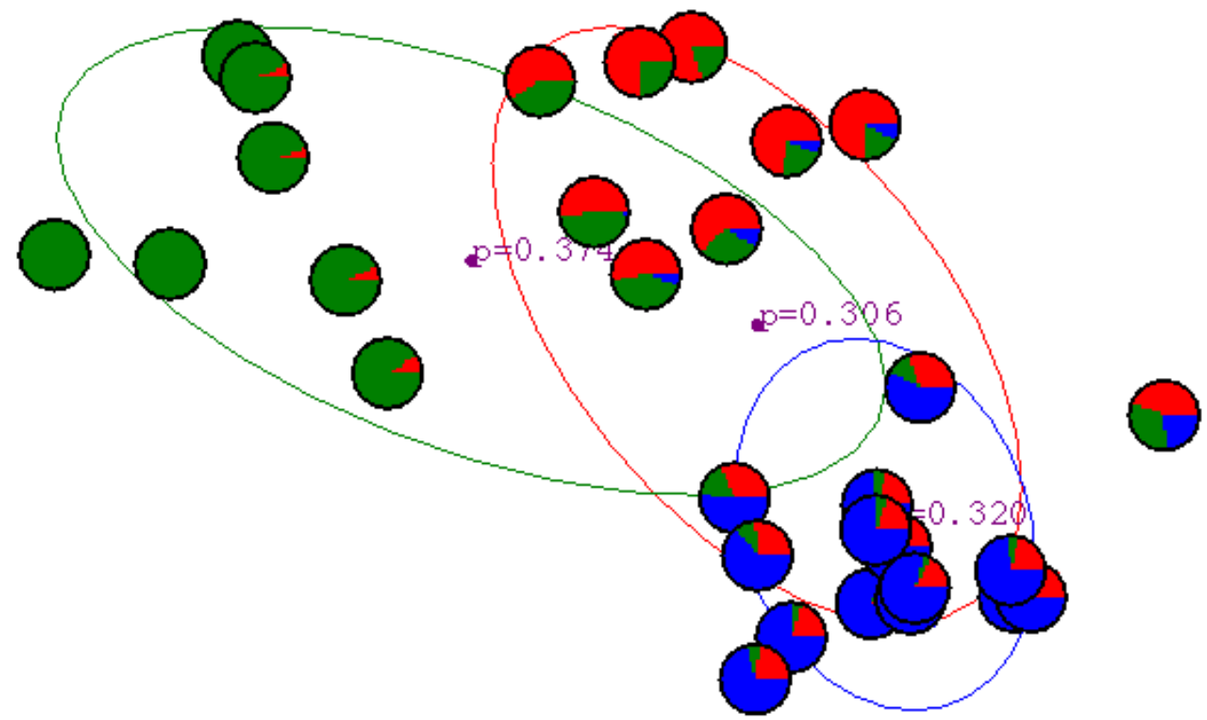
$$\hat{\boldsymbol{\Sigma}}_k^{(h)} := \frac{1}{\sum_{i=1}^{n} w_{ik}^{(h)}} \sum_{i=1}^{n} w_{ik}^{(h)} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(h)})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(h)})^T.$$

After first iteration



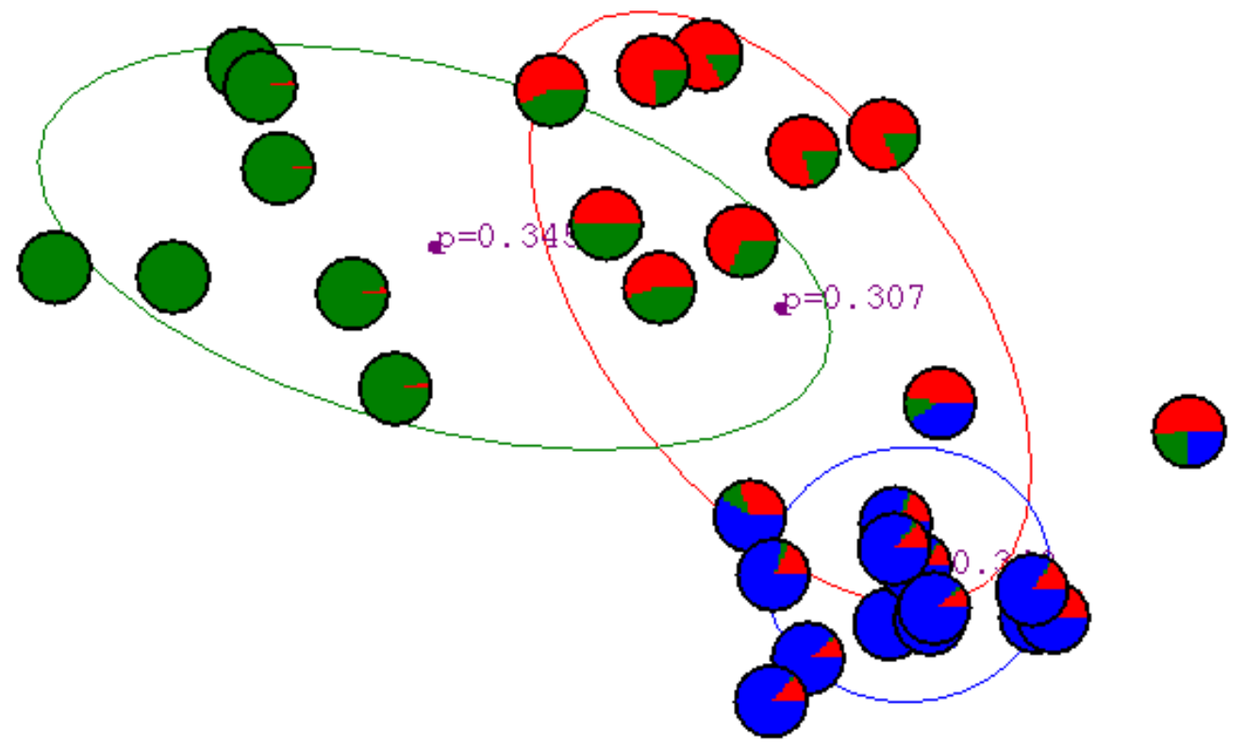
Copyright © 2001, 2004, Andrew W. Moore

After 2nd iteration



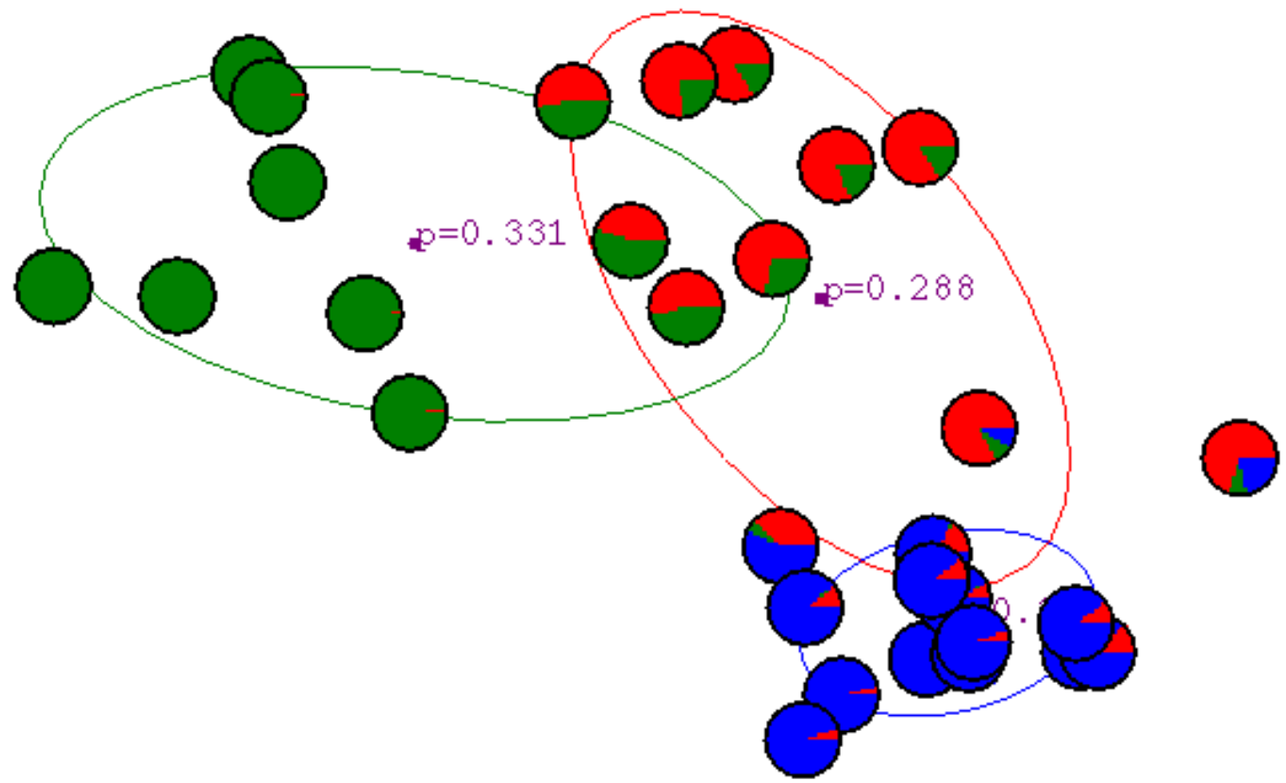
Copyright © 2001, 2004, Andrew W. Moore

After 3rd iteration



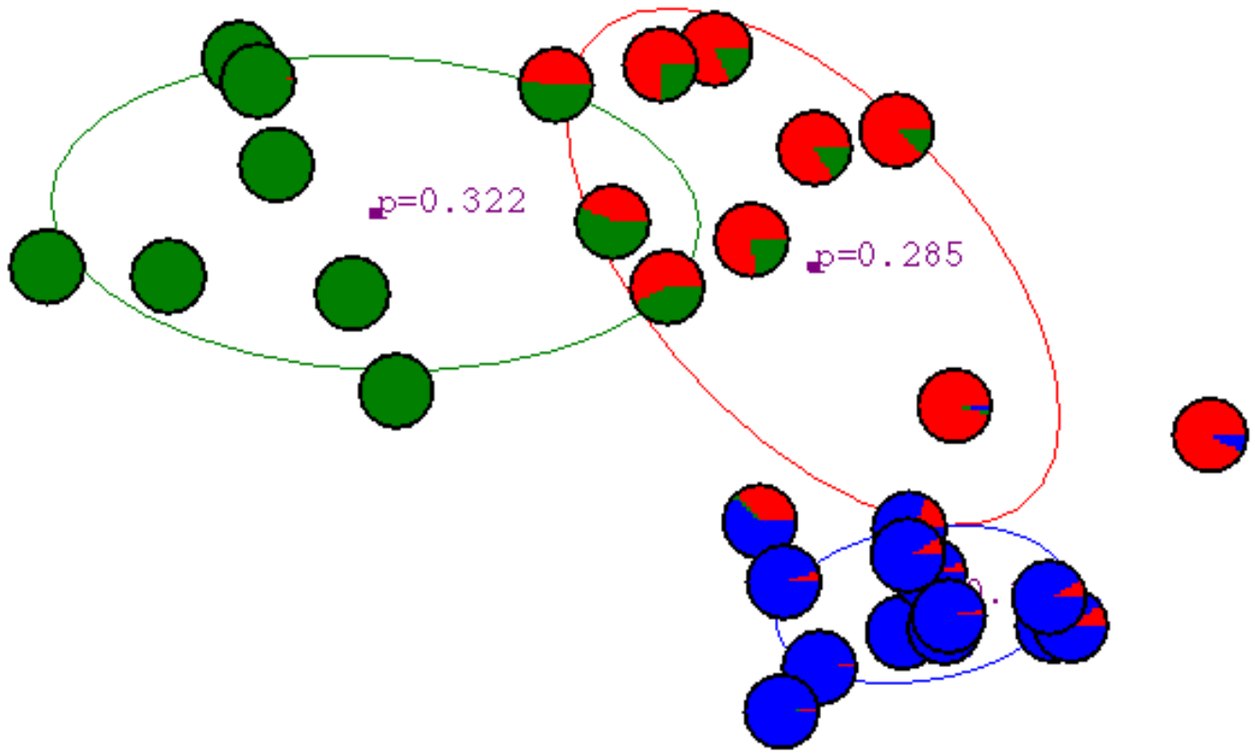
Copyright © 2001, 2004, Andrew W. Moore

After 4th iteration



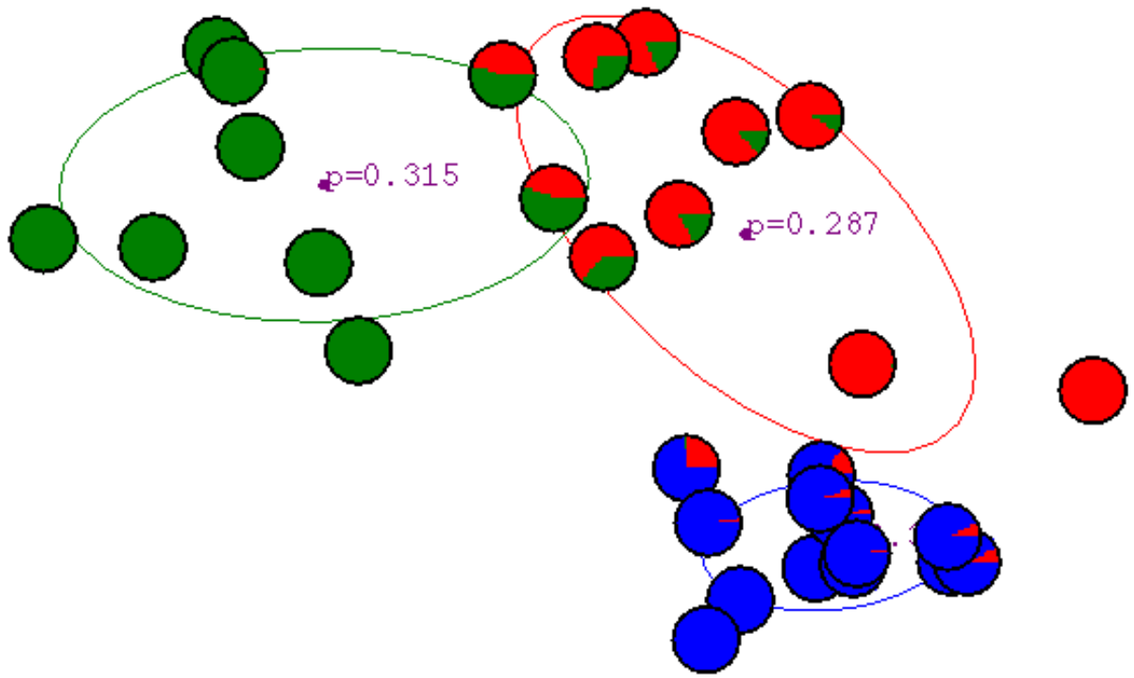
Copyright © 2001, 2004, Andrew W. Moore

After 5th iteration



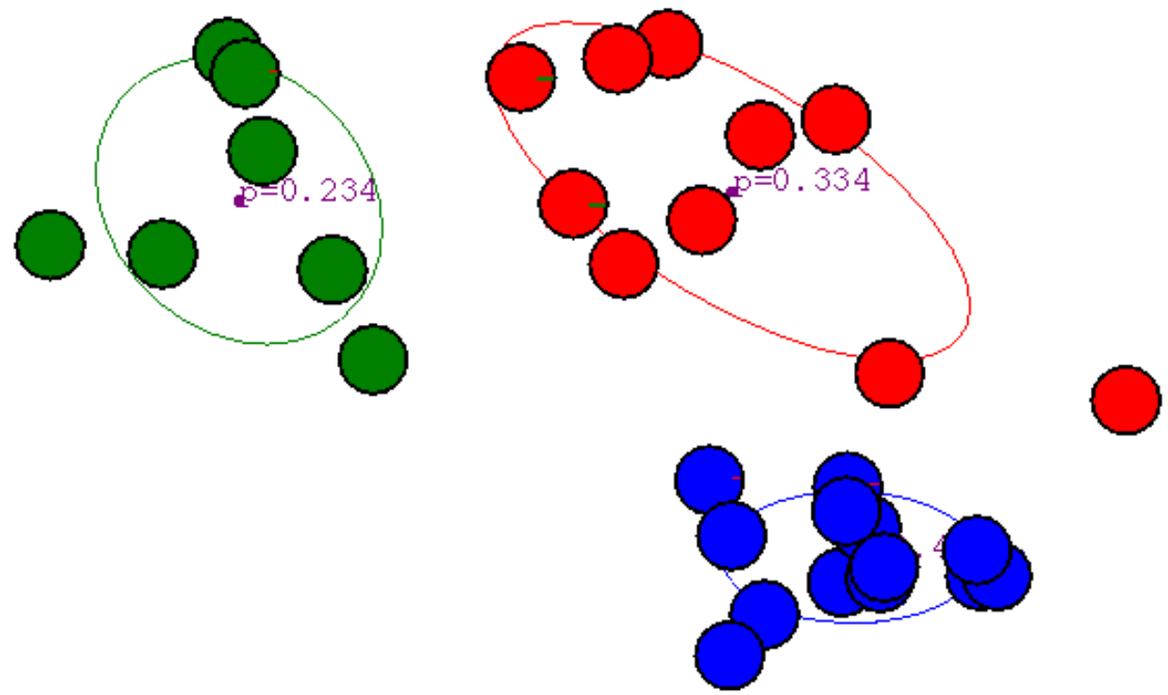
Copyright © 2001, 2004, Andrew W. Moore

After 6th iteration



Copyright © 2001, 2004, Andrew W. Moore

After 20th iteration



Copyright © 2001, 2004, Andrew W. Moore

Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 15

Two-Stage Model-Based Clustering for Liquid Chromatography Mass Spectrometry Data Analysis

Marta Łuksza* Bogusław Kluge† Jerzy Ostrowski‡

Jakub Karczmarski** Anna Gambin††

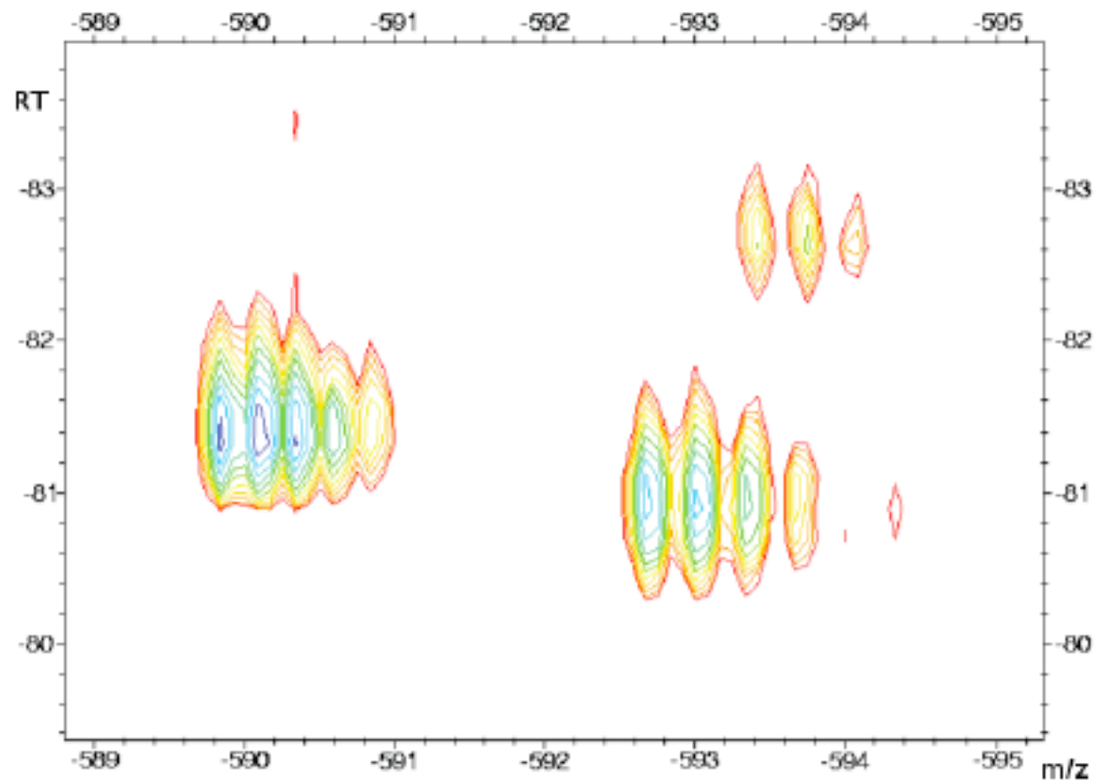


Figure 1: Two-dimensional mass spectrum, a color coded map. The vertical axis is retention time, the horizontal axis is mass-to-charge ratio. Colored spots indicate MS peaks with amplitudes increasing from red to blue. Three peptide signals (isotopic envelopes) are visible.

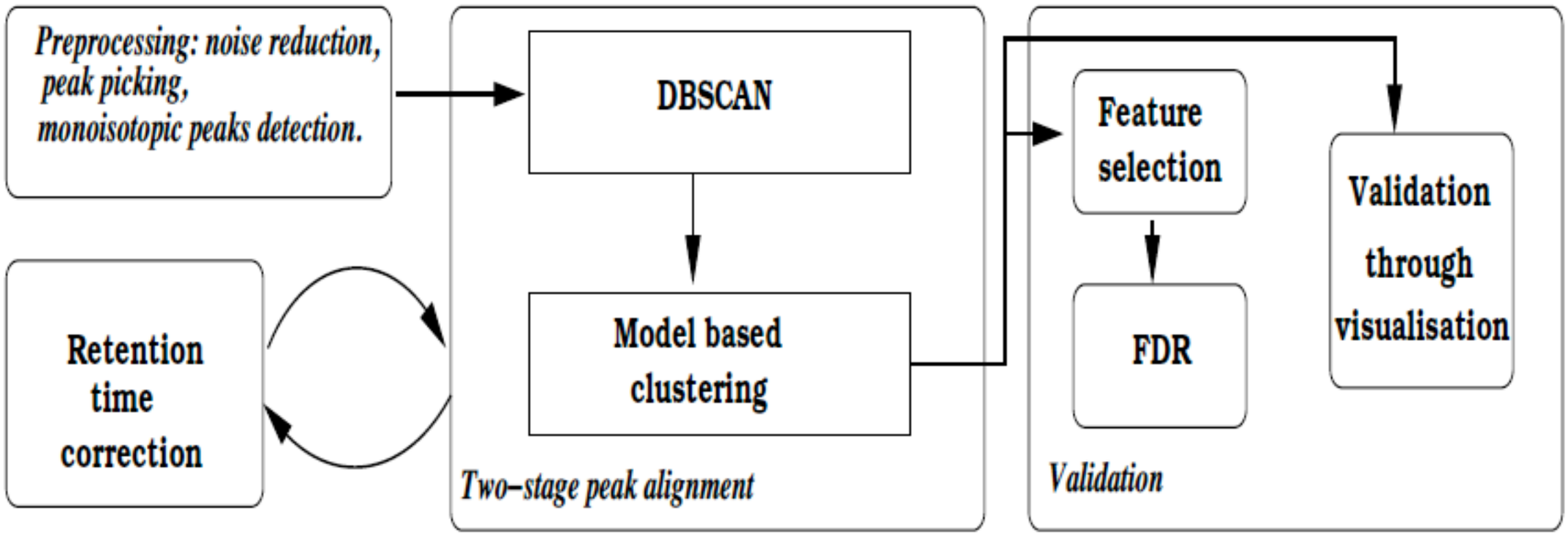


Figure 2: Process flow through different phases of proposed method.

- $[\lambda_k B_k]$ – the most general diagonal model, where the volumes and shapes are allowed to vary between clusters,
- $[\lambda B]$ – a model of identical clusters,
- $[\lambda_k B]$ – a model in which clusters have identical shape but their volumes may vary,
- $[\lambda B_k]$ – a model, where all the clusters have the same volume but their shapes may vary,
- **[fixed variances, equal volume]** – a model with $\Sigma_k = \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \end{pmatrix}$, where $c_1, c_2 > 0$ are constants.
- **[fixed m/z variance, unknown retention time variance]** – $\Sigma_k = \begin{pmatrix} c & 0 \\ 0 & \sigma \end{pmatrix}$, where $c > 0$ is a constant, σ is to be estimated.
- **[fixed m/z variance, unknown varying retention time variance]** – $\Sigma_k = \begin{pmatrix} c & 0 \\ 0 & \sigma_k \end{pmatrix}$, where $c > 0$ is a constant, σ_k is to be estimated for every cluster.

DATA

The data set comprised mass spectra acquired from plasma samples for colorectal cancer patients. Apart from the patient data, control samples were also collected from healthy donors and analyzed with the mass spectrometer. The colorectal cancer data set consisted of 40 spectra, 23 samples corresponding to patients and 17 to healthy donors. Blood samples were collected from patients and age-matched healthy controls. For plasma collection K3 E (greiner bio-one cat.no 455036) tubes supplemented with EDTA were used and after collection samples were centrifuged at 2800g for 15 min at 4°C. Obtained plasma was aliquoted in 200 μl portions, frozen in liquid nitrogen and stored at -70°C for further use. For analysis, plasma aliquots were centrifuged through a 5 kDa or 30 kDa (or both) cutoff filtration membrane (Millipore Ultrafree-MC) in the presence of 20% acetonitrile as a chaotropic agent. Membrane was thoroughly washed with 25% acetonitrile prior to use. To the filtrate the internal standard was added. A HPLC purified peptide (200 pg in each experiment) obtained from tryptic digest of lysozyme (FESNFNTQATNR, mol. mass. 1428.65 Da, m/z = $714{,}84^{+2}$) was used as an internal standard.

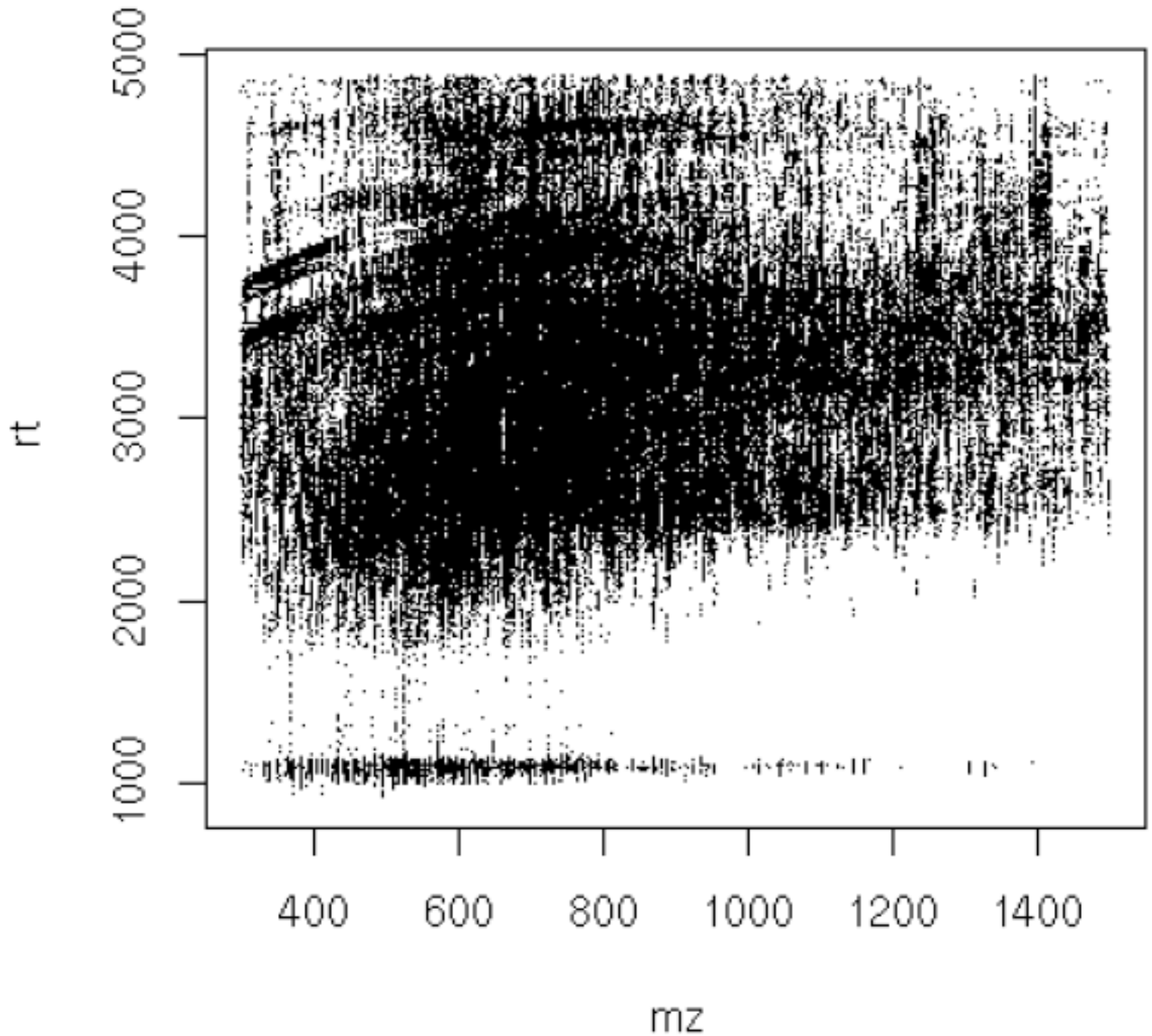


Figure 3: Colorectal cancer data, peaks from 40 samples presented as points in two-dimensional space. Retention time dimension units are seconds.

DBSCAN $\mathcal{O}(n \log n)$

The idea of the algorithm is that for each clustered point the area within the specified radius ε should contain at least $MinPts$ points, where $MinPts$ is a parameter of the algorithm. The algorithm is based on the notion of two points being *density-reachable* with regards to parameters ε and $MinPts$. The ε-neighborhood of point p , $N_\varepsilon(p)$ is the area of radius ε and center at point p. Take two points, p and q. Point q is *directly reachable* from p if q belongs to the ε-neighborhood (i.e. to the area of ε radius) of p. In turn, points p and q are *density-reachable* if there exists a path p_1,..., p_k, where $p = p_1$ and $q = p_q$ such that p_{i+1} is directly reachable from p_i.

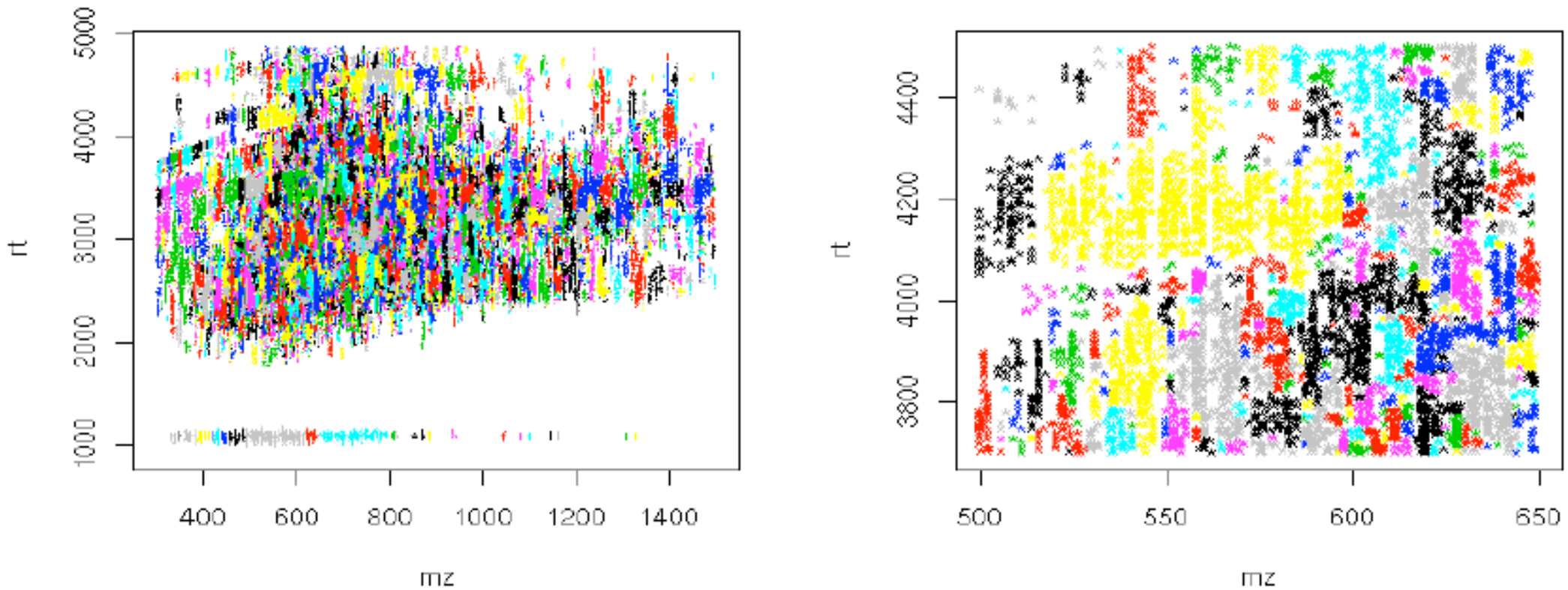


Figure 4: Colorectal cancer data clustered with the DBSCAN algorithm, $\varepsilon_m = 5$, $\varepsilon_{rt} = 30$, $minPts = 10$. Picture on the right presents fragment of the data in greater detail. The cluster colors are recycled.

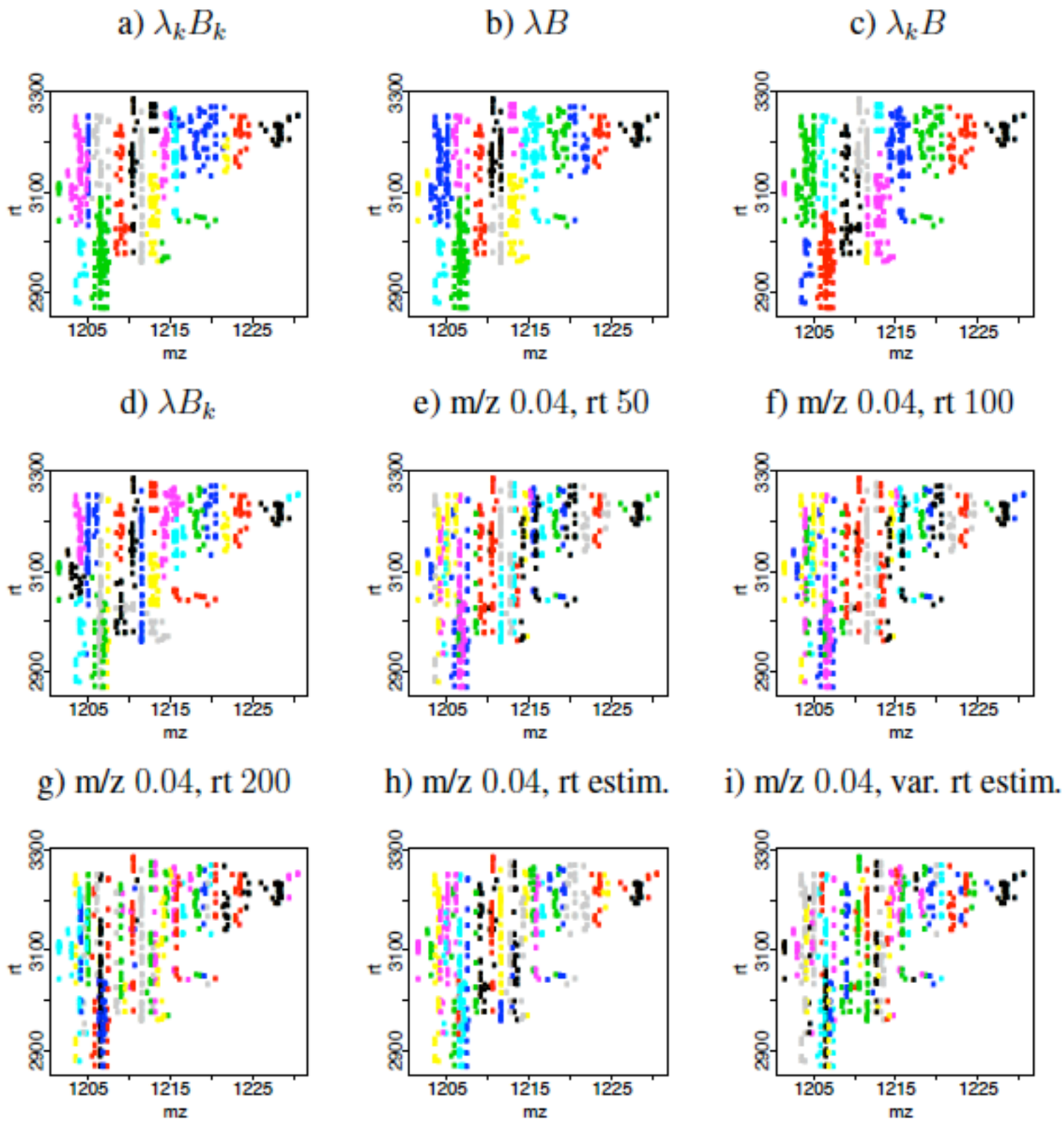
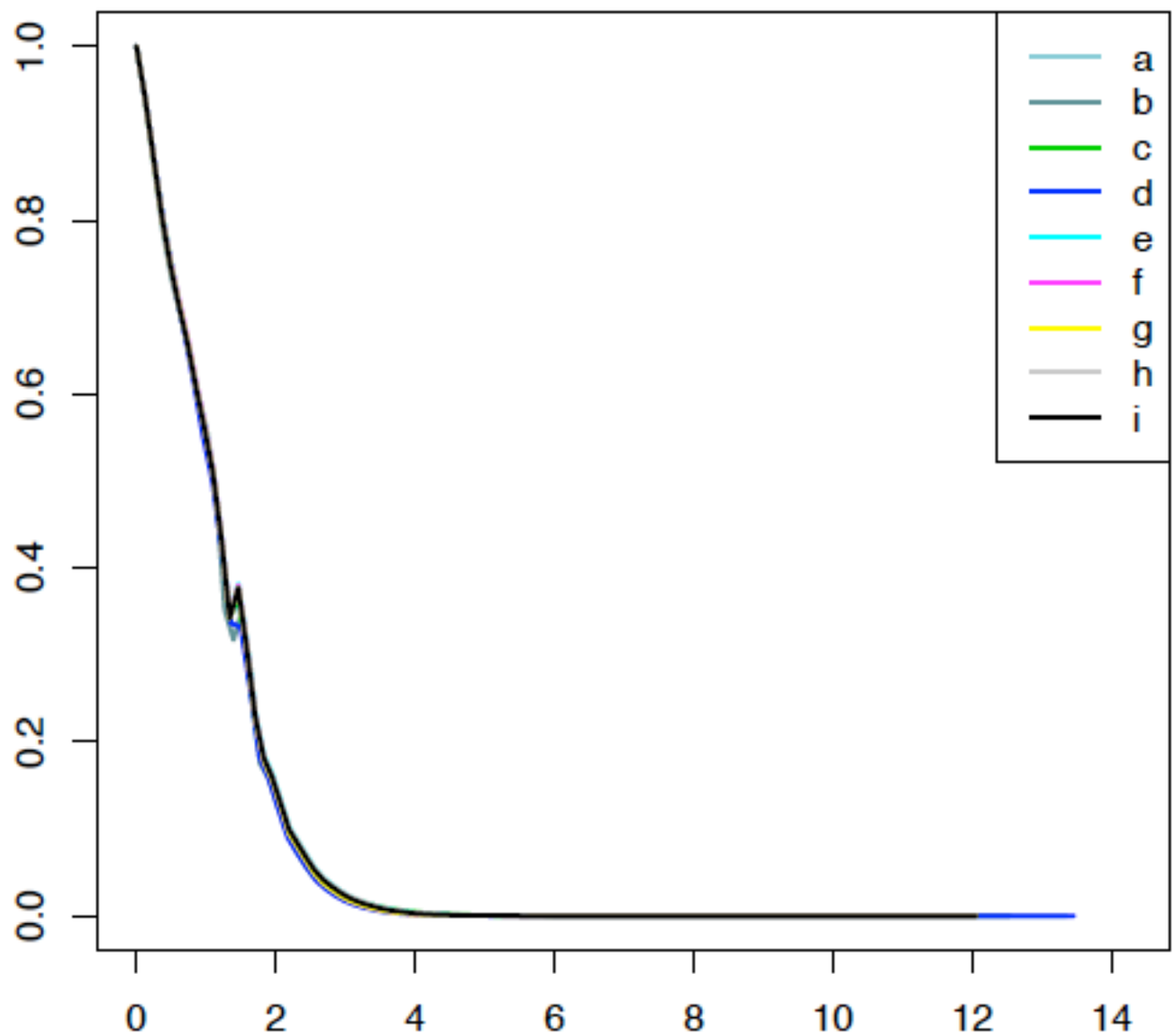


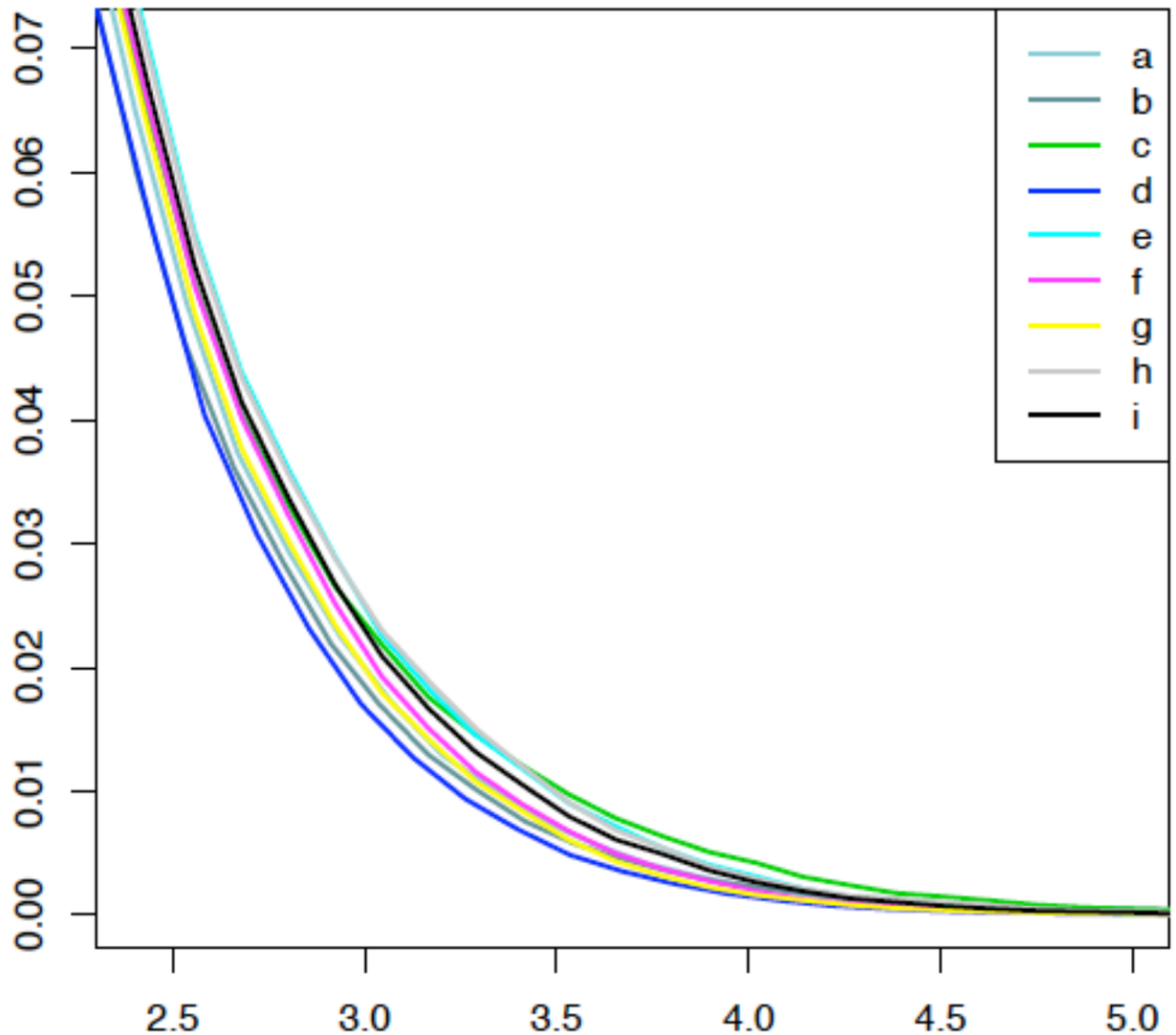
Figure 5: Example of clustering acquired for one of the preliminary clusters from Figure 4 of size 999. a) – i) stand for fitted models: a) $\lambda_k B_k$; b) λB; c) $\lambda_k B$; d) λB_k; e) – i) All the models have deviation in the m/z dimension fixed to 0.04, retention time deviations are: e) 50 f) 100 g) 200 h) estimated from the data, the same for every cluster i) estimated from the data for each cluster.

$$\widehat{FDR}(t) = \frac{\sum_{i=1}^{R} T_i/R}{T}$$

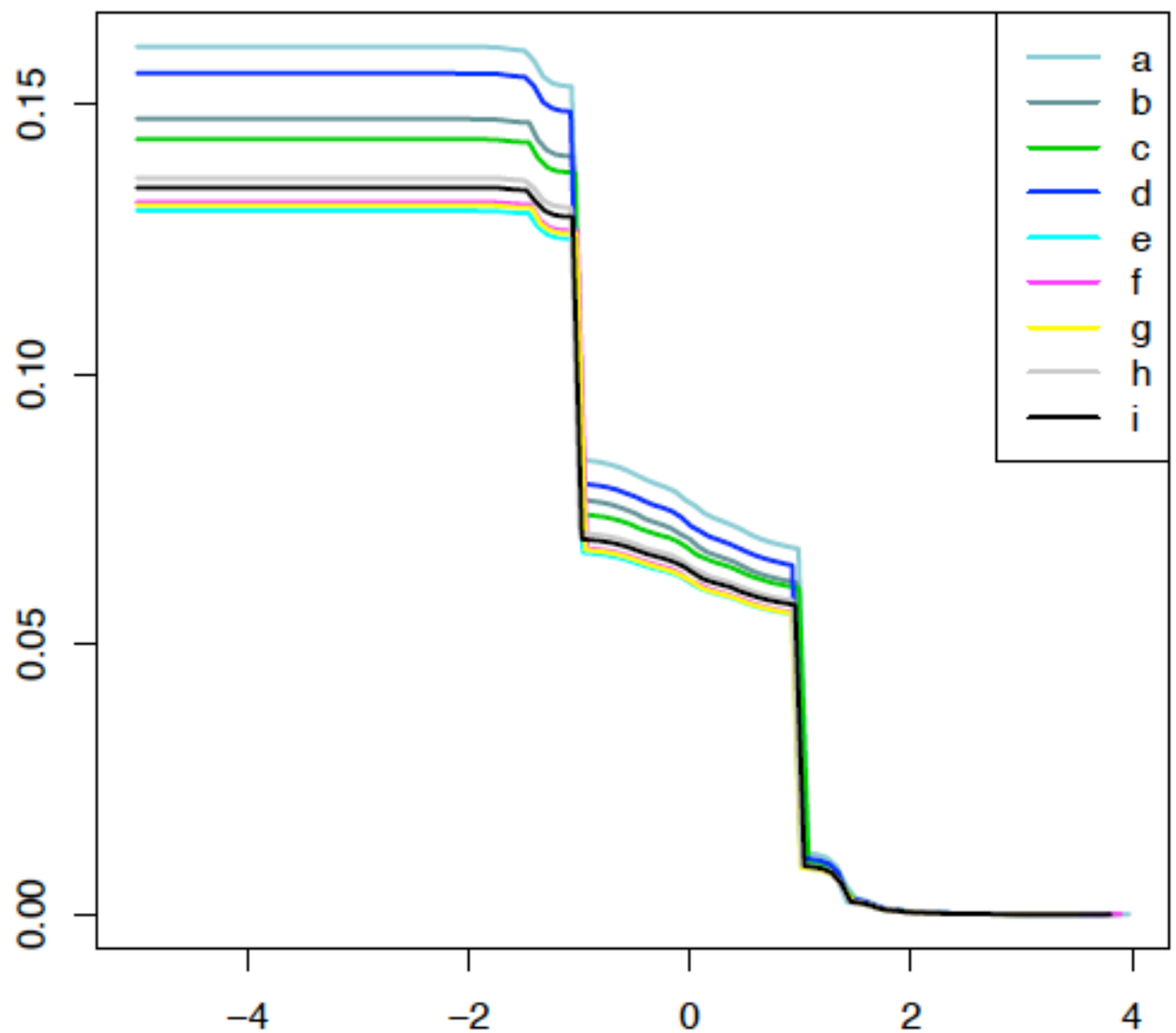
Figure 12: FDR statistic computed for the T-test and Random Forest based feature selection. The vertical axis shows different values of the score assigned to attributes, the horizontal axis shows FDR test values. a) – i) stand for fitted models: a) $\lambda_k B_k$; b) λB; c) $\lambda_k B$; d) λB_k; e) – i) All the models have deviation in the m/z dimension fixed to 0.04, retention time deviations are: e) 50 f) 100 g) 200 h) estimated from the data, the same for every cluster i) estimated from the data for each cluster.

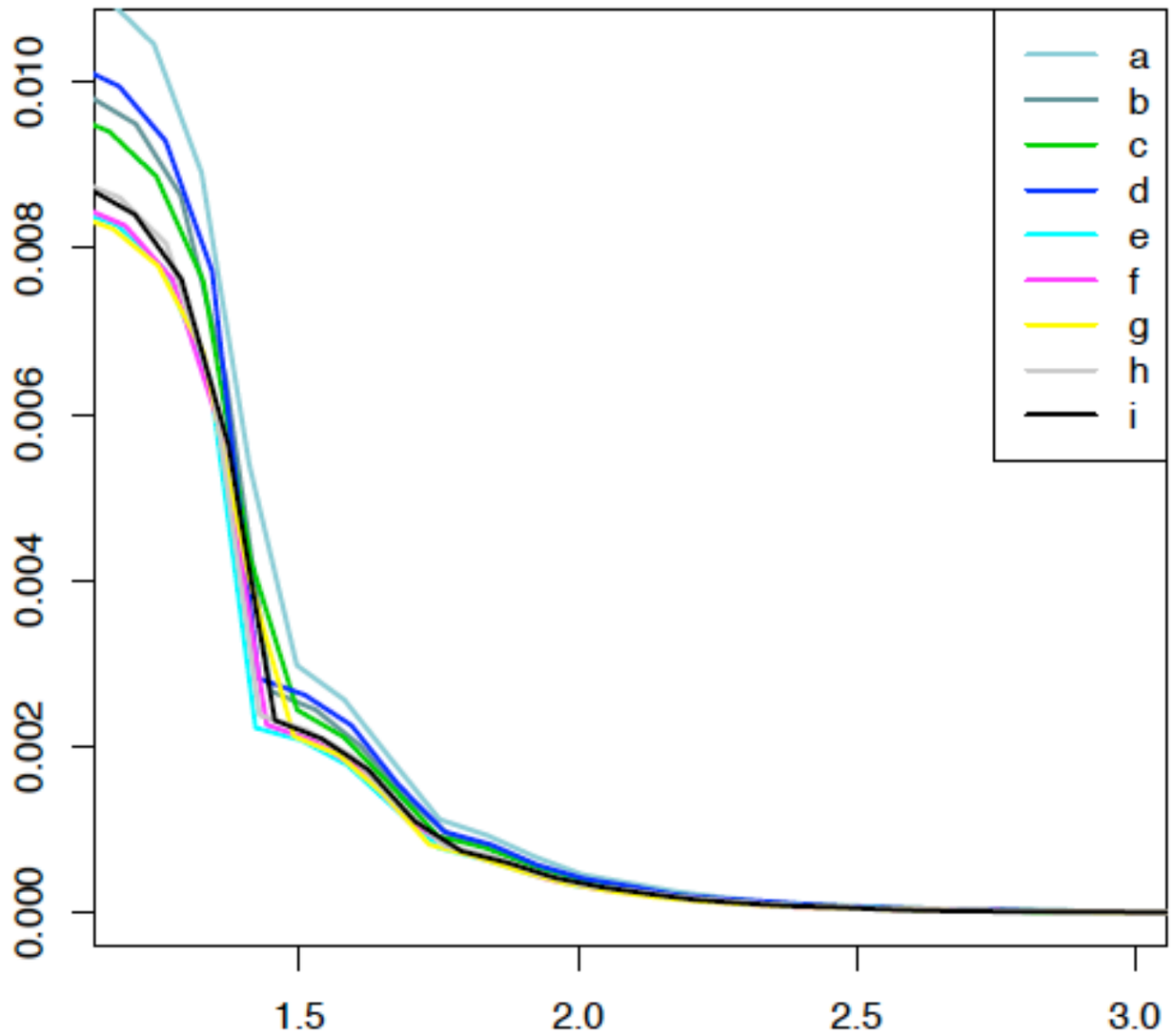
T-test



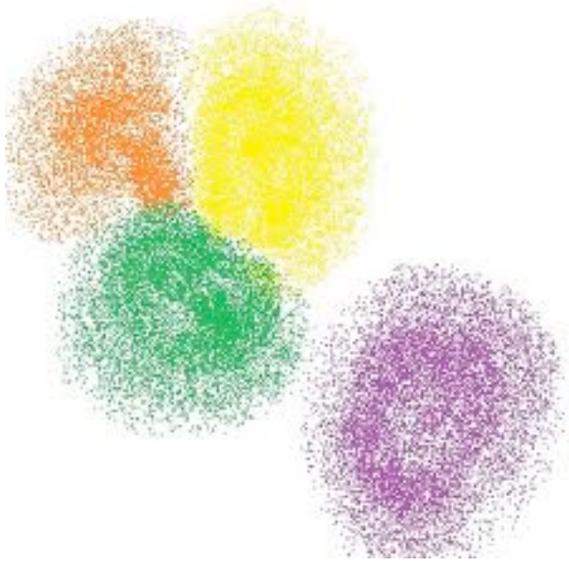

0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
2.5
3.0
3.5
4.0
4.5
5.0
a
b
c
d
e
f
g
h
i

Random Forest




0.000
0.002
0.004
0.006
0.008
0.010
1.5
2.0
2.5
3.0
a
b
c
d
e
f
g
h
i

Doceniając teoretyczną elegancję algorytmu EM musimy odnotować, że jego zbieżność jest dość powolna (tylko liniowa). Kolejną wadą jest, że znajduje on jedynie *lokalne maksimum* funkcji $L(\theta)$ oraz jest bardzo wrażliwy na wybór początkowych wartości $\hat{\pi}_k^{(0)}$, $\hat{\boldsymbol{\mu}}_k^{(0)}$ i $\hat{\boldsymbol{\Sigma}}_k^{(0)}$ dla $k = 1, \ldots K$.



3 Algorytm k-średnich

W modelach z użyciem mieszanin rozkładów normalnych średnie dla grup $\boldsymbol{\mu}_k$ były poprawiane z pomocą ważonych średnich obserwacji x_i. Wagi w_{ik} odpowiadały warunkowym prawdopodobieństwom zdarzenia, że i-ta obserwacja pochodzi z grupy k. Te prawdopodobieństwa były wysokie dla obserwacji leżących w pobliżu środka grupy czyli wartości $\boldsymbol{\mu}_k$.

Na tym pomyśle opiera się algorytm k-średnich [4], który może być postrzegany jako pewne uproszczenie algorytmu EM. Zamiast wyliczać wyrafinowane wagi przypisujemy wszystkie obserwacje do najbliższej grupy, tzn takiej której aktualny środek $\hat{\boldsymbol{\mu}}_k$ położony jest najbliżej. [3] Następnie poprawiamy położenie środków licząc średnią arytmetyczną wszystkich obserwacji przypisanych do danej grupy.

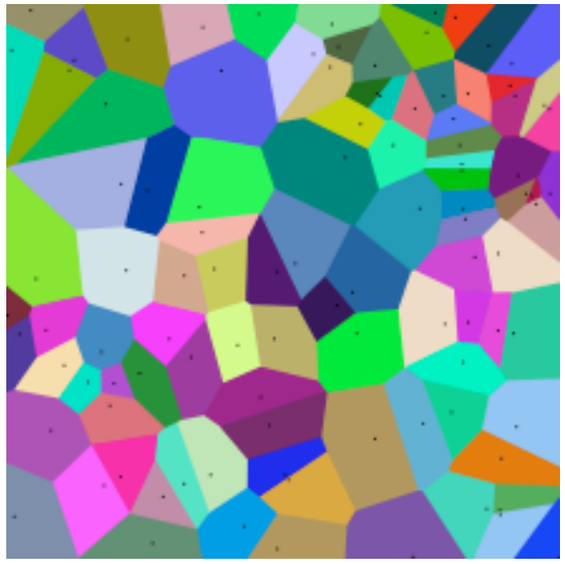
[3] Mówimy cały czas o obserwacjach w $\mathcal{R}^p$ oraz odległości euklidesowej.

Dla odpowiednio dobranych wartości początkowych $\hat{\boldsymbol{\mu}}_1^{(0)}, \ldots, \hat{\boldsymbol{\mu}}_K^{(0)}$ algorytm k-średnich powtarza następujące dwa kroki dla $h = 1, 2, \ldots$:

- Dla każdej obserwacji $\boldsymbol{x}_i$ znajdź najbliższy środek $\hat{\boldsymbol{\mu}}_k^{(h-1)}$ i przypisz tę obserwację do grupy o tym środku.
- Popraw środki grup przez policzenie średniej w każdej grupie:

$$\hat{\boldsymbol{\mu}}_k^{(h)} := \frac{1}{\text{rozmiar grupy } k} \sum_{\text{obserwacje } \boldsymbol{x}_i \text{ należące do grupy } k} \boldsymbol{x}_i$$

Możemy powiedzieć, że algorytm k-średnich znajduje grupy i poprawia ich środki „na sztywno”, w przeciwieństwie do algorytmu EM, który bardziej elastycznie wylicza (warunkowe) prawdopodobieństwo, że dana obserwacja pochodzi z klasy k. Podział przestrzeni jaki otrzymujemy w wyniku algorytmu k-średnich nazywamy regionami Voronoi.



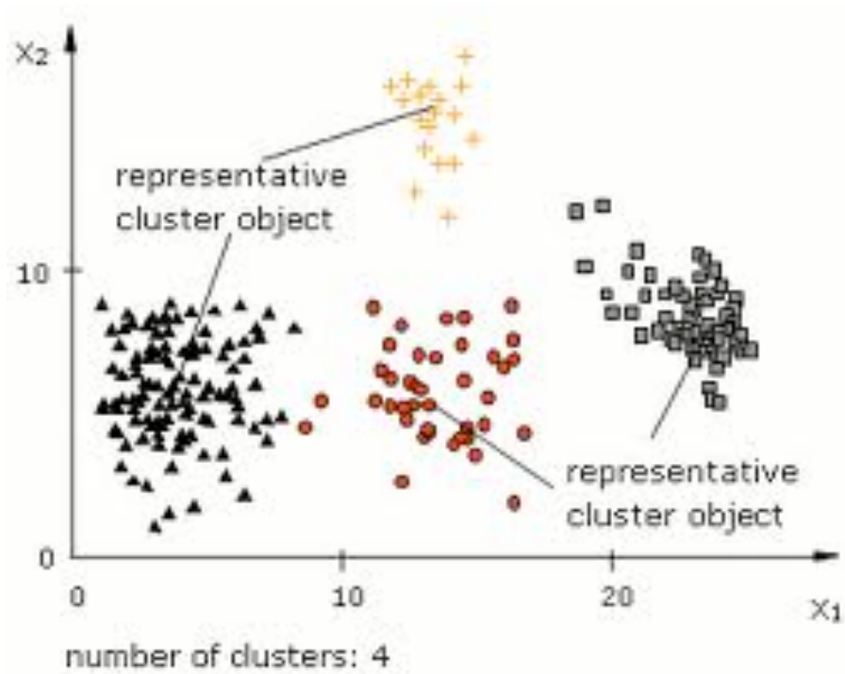
28 April 1868–20 November 1908

Można pokazać, że szybkość zbieżności dla algorytmu k-średnich jest dużo większa niż dla algorytmu EM. Z tego powodu używa się często znalezionego przez k-średnie grupowania, jako początkowych wartości w algorytmie EM.

Podobnie jak algorytm EM, algorytm k-średnich jest bardzo wrażliwy na wybór wartości początkowych. Często stosowaną strategią jest porównanie wielu przebiegów algorytmu dla różnych wartości początkowych. Jeśli za każdym razem uzyskamy inne grupowanie możemy wywnioskować, że danych, z którymi pracujemy, nie daje się w sensowny sposób pogrupować.

4 Algorytm k-median

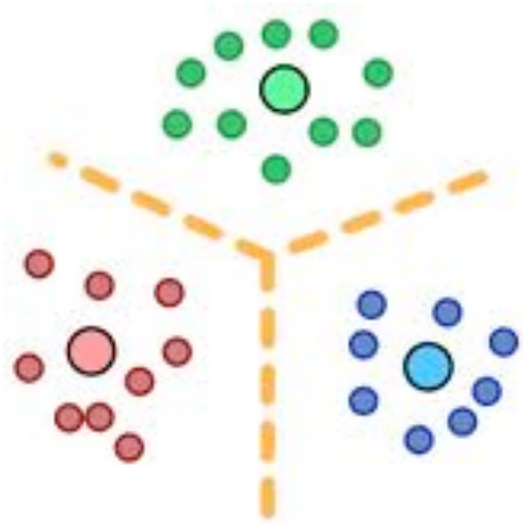
Algorytm k-średnich używa odległości euklidesowej. Jako jego uogólnienie zaproponowano algorytm k-median, który używa dowolnej miary odległości. [4] Często przyjmujemy, że $d_{ij=|\boldsymbol{x}_i-\boldsymbol{x}_j|}$ oraz wymagamy, żeby środki grup były tożsame z pewną obserwacją należącą do danej grupy, czyli $\hat{\boldsymbol{\mu}}_k = \boldsymbol{x}_{c_k}$ dla pewnego c_k.



Algorytm powtarza następujące dwa kroki, aż grupowanie przestanie się zmieniać:

- Każdą obserwację $\boldsymbol{x}_i$ przypisz do grupy k, której środek $\hat{\boldsymbol{\mu}}_k = \boldsymbol{x}_{c_k}$ minimalizuje d_{ic_k}.
- Dla każdej grupy k (gdzie $k = 1, \ldots K$) znajdź centroid, czyli obserwację $\boldsymbol{x}_{c_k}$, która minimalizuje:

$$\sum_{\text{obserwacje } \boldsymbol{x}_i \text{ należące do grupy } k} d_{ic_k}$$



Inicjalizacja algorytmu k-median następuje najczęściej przez wylosowanie indeksów środków grup ze zbioru $\{1, \ldots n\}$ (bez zwracania).

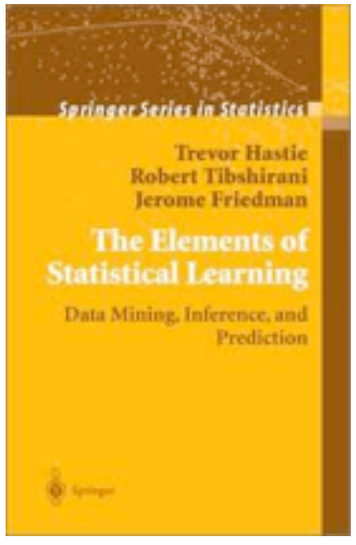
Algorytm k-średnich minimalizował rozrzut obserwacji w grupach mierzony jako:

$$\alpha_K = \sum_{k=1}^{K} \sum_{\text{obserwacje } \boldsymbol{x}_i \text{ należące do grupy } k} ||\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k||^2 \tag{4}$$

Analogicznie algorytm k-median minimalizuje:

$$\sum_{k=1}^{K} \sum_{\text{obserwacje } \boldsymbol{x}_i \text{ należące do grupy } k} d_{ic_k}$$

Niestety obydwa algorytmy nie gwarantują zbieżności do wartości minimalnej.

Zastanówmy się jaszcze jak dobierać liczbę grup K. Narzucające się rozwiązanie to wizualizacja danych, o ile są „niskowymiarowe”. Jeżeli dane, które analizujemy, mają wysoki wymiar możemy zastosować skalowanie wielowymiarowe, które będzie przez nas omawiane niebawem, żeby ten wymiar zmniejszyć. Niektórzy autorzy proponują zastosowac techniki analizy dyskryminacyjnej, w celu uzyskania podziałów na grupy.

Wspomnijmy jeszcze o zaproponowanym przez Tibshiraniego, Waltera i Hastie podejściu zwanym **statystyką odstępu**. Pomysł polega na tym żeby obliczać wartość α_k ze wzoru (4) dla kolejnych $K = 1, 2, \ldots$ i wybierać taką wartość, dla której różnica pomiędzy kolejnymi α_K jest minimalna.