Uniwersytet Warszawski
Wydział Filozofii i Socjologii. Instytut Socjologii.

Piotr Zimolzak
Nr albumu: 219468

Analiza skupień jako szczególny przypadek skalowania

Praca magisterska
na kierunku SOCJOLOGIA

Praca wykonana pod kierunkiem
dr. Henryka Banaszaka
Zakład Statystyki, Demografii i Socjologii Matematycznej

Maj 2009

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

Analiza skupień jest zbiorem metod prowadzących do podziału obiektów. Od takiego podziału wymaga się, aby jego elementami były klasy maksymalnie homogeniczne wewnątrz i maksymalnie heterogeniczne między sobą. Analiza skupień jest obecnie potężnym narzędziem w rękach analityka. Zaskakującym jest, że mimo długiej tradycji badań metodologicznych, nie doczekała się ona opracowania na temat jej podstawowej problematyki.

Przez wiele lat analiza skupień rozwijała się i rozwija w ramach wielu dyscyplin. Wraz z upływem czasu, zaczęto dostrzegać jej podstawowe braki zarówno od strony metodologicznej jak i algorytmicznej. Zaczęły pojawiać się mniej lub bardziej wyrafinowane usprawnienia w postaci różnych kryteriów, wskaźników, współczynników zdających sprawę z jakości uzyskanych podziałów.

Jednak sposób myślenia o analizie skupień nie zmienił się. Do przełomu lat siedemdziesiątych i osiemdziesiątych ubiegłego stulecia nadal myślano o niej jako o metodzie eksploracyjnej wykorzystywanej *ad hoc* do rozwiązania konkretnych problemów.

Przełom nastąpił we wspomnianym wyżej okresie. Pojawiła się wówczas próba opisania analizy skupień w języku modelowania statystycznego. W nowej koncepcji zaczęto wykorzystywać wiedzę na temat estymacji i wnioskowania statystycznego. Początkowo rozwój metodologii przebiegał zgodnie z dwoma niezależnymi nurtami: analizą profili ukrytych i modelami mieszanymi. Podobieństwo między dwiema perspektywami zostało zauważone przez Wolfe'a na początku lat siedemdziesiątych, a formalnej syntezy dokonał L. Goodman w 1974 roku. W jego pracy pojawiła się iteracyjna metoda estymacji, która w roku 1977 została sformalizowana w klasycznej pracy Dempstera i in., której nadano nazwę algorytmu EM. Obecnie jest on podstawowym elementem wielu pakietów statystycznych do analizy struktury ukrytej.

Celem niniejszej pracy jest próba odpowiedzi na pytanie, czy istnieje metoda analizy skupień gwarantująca uzyskanie „dobrego podziału" zbioru obiektów. Zostaną sformułowane postulaty pod adresem metody, która potrafiłaby w uzasadniony sposób wyjaśnić, co kryje się pod pojęciem „dobrego podziału". Zostanie pokazane, że klasyczna (w sensie sposobu patrzenia na metodologię) analiza skupień nie jest w stanie sprostać stawianym postulatom.

Zmiana perspektywy oznacza sprowadzenie problemu analizy skupień do zagadnienia skalowania. Skalowanym obiektem jest pewna cecha ukryta - klasyfikacja, którą jest skalowana za pomocą obserwowalnych wskaźników - zmiennych biorących udział w procesie analizy skupień. Model skalowania posiada przejrzysty opis metodologiczny, dzięki czemu dysponujemy konkretną listą postulatów dobrej metody. Pokażemy, że warunkiem dobrego uzasadnienia generowanego podziału jest przyjęcie perspektywy probabilistycznej. W ostatniej części pracy

zostaną przedstawione problemy związane z testowaniem hipotez na temat struktury zbioru, a na koniec zostaną zilustrowane wybrane metody probabilistyczne wykorzystywane w popularnych pakietach statystycznych.

Słowa kluczowe

analiza skupień, skalowanie, cecha ukryta

Dziedzina pracy (kody wg programu Socrates-Erasmus)

Statystyka

Klasyfikacja tematyczna

Statystyka, data-mining

Tytuł pracy w języku angielskim

Cluster Analysis as a Specific Example of Scaling a Latent Trait

Spis treści

Stosowane oznaczenia

W niniejszej pracy zostanie wykorzystany następujący system oznaczeń:

Tabela 1: Podstawowe oznaczenia wykorzystywane w pracy

X	zbiór obserwacji (obiektów)
n	liczba obserwacji w zbiorowości (jeśli nie zaznaczymy inaczej, chodzić będzie o próbę)
p	wymiar przestrzeni (liczba zmiennych opisujących obserwacje)
x_i	profil obserwacji o numerze i (p-wymiarowy wektor)
x_i^m	wartość profilu na współrzędnej o numerze m (wartość zmiennej losowej na miejscu m)
$f(x)$	gęstość uzyskania danego profilu
C_i	skupienie (klasa) o numerze i
k	liczba skupień
λ_j	częstość (proporcja) występowania skupienia (klasy) o numerze j
θ_j	wektor parametrów opisujących skupienie (klasę) o numerze j
μ	wartość oczekiwana zmiennej o (wielowymiarowym)rozkładzie normalnym
σ	odchylenie standardowe zmiennej o (wielowymiarowym)rozkładzie normalnym
$\hat{\theta}$	estymator wektora parametrów (w ogólnej postaci)
$f(x_i\|\theta_j)$	warunkowa gęstość profilu i pod warunkiem przynależności do skupienia j
$d_{i,j}$	odległość między obserwacjami i, j
$\mathcal{R}$	relacja dwuargumentowa
$\mathcal{B}_k$	podział zbioru obserwacji na k bloków
$\mathcal{A}$	przykładowy zbiór amorficzny (zob. Aneks)
$\mathcal{S}$	przykładowy zbiór skrystalizowany (zob. Aneks)

Rozdział 1

Wprowadzenie

1.1. Analiza skupień jako metoda podziału zbioru obiektów

Analiza skupień należy do szerokiej rodziny metod zajmujących się efektywnym podziałem zbioru obiektów na grupy. Metody otrzymywania takiego podziału są różne, bo różne są obiekty i różne ich cechy brane są pod uwagę. Tym, co łączy różne metody jest dążenie do tego, aby obiekty stanowiące jedną grupę były w maksymalnym stopniu jednorodne (homogeniczne), a wyróżnione grupy maksymalnie różne (heterogeniczne) między sobą.

1.1.1. Definicja. **Analizą skupień** będziemy nazywać zbiór metod i algorytmów wyznaczania podziału zbioru obiektów na podzbiory przy wykorzystaniu wyłącznie informacji na temat cech każdego obiektu, tak aby utworzone podzbiory były możliwie najbardziej jednorodne, rozłączne i wyczerpywały cały zbiór obiektów.

Reformułując nadrzędne zadanie analizy skupień, możemy powiedzieć, że ogólnym celem analizy skupień jest taki podział zbioru obiektów na grupy, aby **różnice między grupami były większe niż między obiektami wewnątrz grup**.

W przeciwieństwie do wielu znanych nam metod analizy danych, analiza skupień jest raczej **zbiorem metodologii**, które rozwijały się poza poza obrębem jednej dyscypliny. Korzyścią płynącą z takiego obrotu rzeczy był dynamiczny rozwój metodologii. Jednak z drugiej strony, równoległe i nieskoordynowane działania nieraz prowadziły do podobnych rozwiązań, ale ukrytych pod różnymi nazwami. To zróżnicowanie nazw dla, identycznych w gruncie rzeczy metod i wyników, wielokrotnie uniemożliwiało skuteczne ich porównywanie.

Analiza skupień funkcjonuje obecnie pod wieloma synonimicznymi pojęciami. Do najbardziej znanych należą: grupowanie, segmentacja, taksonomia numeryczna, klasyfikacja taksonomiczna lub uczenie się bez nadzoru (ang. unsupervised learning). Zatem bardziej kluczową kwestią jest doprecyzowanie modelu niż stosowanie danej terminologii.

1.1.1. Metody analizy skupień a jej algorytmy

W dalszym ciągu będziemy rozróżniać pojęcia **metody** i **algorytmu**. Niezbędnym elementem metody jest pewien model lub założenie na temat „natury" rzeczywistości. Przykładem takiego modelu jest założenie na temat losowego charakteru zmiennych opisujących daną zbiorowość. W ramach metody definiowane są relacje między każdą parą obiektów. Te z kolei przekładane są na pewne miary podobieństwa między nimi. Przykładem jest tutaj macierz odległości między obiektami. W tym miejscu pojawia się ważny element metody, czyli algorytm. Jest zbiorem instrukcji, który w skończonej liczbie kroków przetwarza informację na temat miary podobieństwa w podział zbioru obiektów zgodny z definicją analizy skupień.

Metoda jest pojęciem znacznie szerszym i może obejść bez algorytmu, podczas gdy relacja odwrotna nie jest zachowana. Na metodę będzie składać się zarówno założenie na temat budowy rzeczywistości (model) jak i pomysł, jak w ramach tej koncepcji zrealizować cel. Z kolei algorytm będziemy rozumieć jako pewien przepis lub zbiór instrukcji, który w skończonej liczbie kroków pozwala (lub nie) osiągnąć postulat wynikający z metody.

Jak czytamy w [47, s.16], metody skupiania nie zakładają żadnego kryterium przypisywania obiektów do grup. Jedynym wewnętrznym kryterium podziału podziału jest matematycznie zdefiniowane podobieństwo między obiektami. Zupełna dowolność modelu z jednej strony pozwala na swobodę w doborze metod, z drugiej zaś prowadzi do niejednoznaczności rozwiązań. Nie ulega wątpliwości, że w analizie skupień drzemie olbrzymi potencjał, jeśli chodzi o podejmowanie ważnych decyzji. Niemniej jednak, wykorzystanie tego potencjału, zależy od wielu rozważnych kroków podjętych przez przystąpieniem do analiz.

1.2. Rys historyczny metodologii

Praktyczne przykłady

Praktycznych zastosowań analizy skupień można by podać bardzo wiele. Dlatego poniższy chronologiczny przegląd zastosowań analizy skupień 1.1 ograniczymy do zaznaczenia przełomowych prac, jeśli chodzi o rozwój metodologii.

Interesującym jest, że socjologowie zaczęli korzystać z analizy skupień relatywnie późno. Pierwsze nawiązania widać w pracy Lazarsfelda i Henry'ego z roku 1968 na temat bardziej ogólnego zagadnienia *analizy ukrytej struktury*. Rozwój analizy skupień na przełomie lat sześćdziesiątych i siedemdziesiątych związany był z zagadnienia skalowania wielowymiarowego.

Szybki rozwój technologii i mocy obliczeniowej komputerów ciągu ostatniej dekady, a w konsekwencji zwiększenie precyzji arytmetycznej pozwoliło na szersze wykorzystanie dokładnych obliczeń niezbędnych dla bardzo małych liczb (pojawiających się np. w estymacji metodą największej wiarogodności). Analiza skupień jest również spoiwem między wieloma dziedzinami, które rozwijały się jak dotąd w separacji. Mam tu na myśli sieci neuronowe, eksploracyjną analizę wzorców, kompresję danych czy analizę obrazów zob. [35, s.269, s.307-310]

Analiza skupień jest również nieocenionym narzędziem w ręku badacza marketingowego (zob. [17, s. 827]). Najczęściej służy ona do przeprowadzenia segmentacji klientów, użytkowników, produktów, a nawet całych rynków. W wielu sytuacjach stanowi jedynie podstawę do dalszych analiz. Warto zauważyć, że bazując na dużych próbach jest dość kosztownym przedsięwzięciem, przez co przeprowadzana jest znacznie rzadziej niż inne metody. Niemniej jednak, jak zauważają Punj i Steward [52, s.135] ostateczne wykorzystanie danej metody jest wypadkową jej popularności, dostępności w komercyjnym oprogramowaniu oraz kosztu z tytułu nakładu pracy i czasu. Kwestia poprawności metodologicznej często schodzi na dalszy plan. Dlatego studiowania analizy skupień może dostarczyć cennej wiedzy i praktycznych wskazówek czyniąc cały proces analityczny bardziej efektywnym.

Można znaleźć również inne ciekawe zastosowania analizy skupień w naukach społecznych. W [35, s.312] czytamy, że Wulfekuhler i Punch wykorzystali ją w 1997 roku do ilościowej analizy treści identyfikując często współwystępujące trzony słów na wybranych stronach WWW. Co raz częściej z tych metod korzystają serwisy internetowe, oferując streszczenie zawartości witryny w postaci tzw. chmur słownych (ang. word clouds lub tag clouds). [1] W najprostszej postaci chmury biorą pod uwagę tylko częstość występowania danego słowa. Dlatego bardziej interesujące z punktu widzenia analizy skupień są tzw. *colocate clouds*, które

[1] Przykładowe chmury można obejrzeć np. na http://www.wordle.net/gallery.

Tabela 1.1: Metodologiczna ewolucja analizy skupień

Rok	Autorzy	Dziedzina	Opis
1791	P.Camper	antropometria	grupowanie szczątków na postawie wymiarów twarzy
1901	K.Pearson	statystyka	metoda doboru płaszczyzn do układu punktów w przestrzeni wielowymiarowej
1911	J.Czekanowski	statystyka	konstrukcja współczynnika bliskości (dziś znanego pod nazwą uśrednionej metryki miejskiej)
1913	J.Czekanowski	antropologia	klasyfikacja 47 plemion afrykańskich na podstawie 17 cech kultury materialnej
1939	Tryon	statystyka	rozwinięcie metod niehierarchicznych, autor pojęcia "analizy skupień"
1951	K.Florek, J. Łukasiewicz, J.Perkal, H.Steinhaus	statystyka	stworzenie metody dendrytowej, znanej jako taksonomia wrocławska
1954-1967	McQuitty	psychologia	rozwinięcie szerokiej klasy metod, które znane są jako analiza wzorców (ang. pattern analysis)
1957-1971	Coleman, Bloombauw	statystyka	zagadnienia redukcji reprezentacji danych w niskowymiarowej przestrzeni (ang. smallest space analysis) i skalowania wielowymiarowego
1963	Ward	statystyka	odkrycie i zastosowanie metody Warda do metod hierarchicznych opartej na analizie wariancji
1967	Johnson	statystyka	formalizacja aglomeracyjnych metod hierarcznych oraz implementacja w języku FORTRAN
1967	McQueen	statystyka	opisanie i zaimplementowanie algorytmu K-średnich (ang. K-means)
1967	Bertin Hartigan	statystyka	reprezentacja zbioru danych za pomocą dendrogramów
1968	Lazarsfeld, Henry	socjologia	opracowanie analizy struktury ukrytej (ang. latent structure analysis), która później stała się impulsem do budowy modeli statystycznych w analizie skupień
1969	Ruspini	statystyka	wykorzystanie teorii zbiorów rozmytych (ang. fuzzy sets theory) w analizie skupień
1970-1971	Wolfe	statystyka	wykazanie analogii między analizą klasy ukrytej a modelami mieszanymi (ang. mixutre models), wyprowadzenie wzoru na estymatory NW parametrów rozkładu

Tabela 1.2: Metodologiczna ewolucja analizy skupień - c.d.

Rok	Autorzy	Dziedzina	Opis
1974	L.Goodman	statystyka	opracowanie iteracyjnej procedury estymacji parametrów dla modeli z cechą ukrytą, wyznaczenie warunków na jednoznaczność rozwiązania
1977	Dempster, Rubin, Laird	statystyka, analiza numeryczna	opisanie i zaimplementowanie algorytmu EM (ang. expectation-maximalization), na którym bazują prawie wszystkie pakiety do modelowej analizy skupień
1981	P.Arabie	statystyka	reprezentacja klasyfikacji rozmytej za pomocą zazębiających się skupień (ang. overlapping clusters)
1984	Kohonen	sieci neuronowe	wykorzystanie sieci neuronowych i uczenia maszyn do analizy skupień w oparciu o samo organizujące się mapy obiektów (ang. SOM - self-organising maps)

uwzględniają współwystępowanie danych słów. Jeszcze bardziej zaawansowaną metodologią są samo-organizujące się mapy pojęciowe (ang. self-organizing maps). Często wizualizuje się je za pomocą tzw. mapy ciepła (ang. heatmap), która jest szczególnym przypadkiem histogramu na płaszczyźnie. Przedstawia ona różne barwy w zależności od natężenia danej cechy (np. częstości występowania, gęstości, odległości). Z reguły, większe natężenie reprezentowane jest przez cieplejsze kolory. [2] Służy ona przede wszystkim do wstępnej naocznej analizy struktury zbioru.

1.3. Model analizy skupień

Przedstawiony poniżej aparat matematyczny jest zbyt ubogi do opisu całego zakresu metodologii analizy skupień. Warto go jednak wprowadzić, aby umieć powiązać intuicyjne pojęcia analizy skupień ze ścisłymi obiektami matematycznymi. Aby ułatwić ich bezpośrednie przełożenie na rzeczywistość, po każdej wprowadzonej definicji terminu będzie przedstawiony jej odpowiednik w analizie skupień.

1.3.1. Definicje

Zacznijmy od przytoczenia podstawowych definicji z teorii mnogości, które będziemy systematycznie przekładać na grunt analizy skupień. Poniższe sformułowania i symbole zostały zapożyczone z [31].

1.3.1. Definicja. **Relacją dwuargumentową** $\mathcal{R}$ nazywamy dowolny zbiór, którego elementami są wyłącznie pary uporządkowane. Mówimy, że x **jest w relacji** $\mathcal{R}$ z y lub między elementami x, y **zachodzi relacja** $\mathcal{R}$, co w obu przypadkach zapisujemy $x\mathcal{R}y$.

Przykładem relacji jest odległość między obiektami. Ten rodzaj relacji jest jednak mało interesujący, gdyż (poza szczególnymi przypadkami), wszystkie pary obiektów są w relacji

[2] Przykład takiej mapy dla analizy treści grupy dyskusyjnej można znaleźć w [32, s.486].

odległości (relacja jest spójna). Jeśli jednak dołączymy do relacji określenie *bliski* to zauważymy, że niekoniecznie wszystkie obiekty muszą być ze sobą w **relacji bliskiej odległości**. Relacja ta ma ponadto szereg ciekawych własności: jest **zwrotna** (bo każdy obiekt jest blisko samego siebie), **symetryczna** (z własności odległości) i **przechodnia** (jeśli obiekt A jest blisko B, a B blisko C, to A nie powinien być daleko od C). Ta ostatnia własność może budzić pewne wątpliwości, ponieważ łatwo możemy wyobrazić sobie sytuację, gdy obiekty A_1 i A_n są w rzeczywistości bardzo odległe, ale na skutek pojedynczych porównań między A_j i A_k, $j, k = 2, 3, \ldots, n-1$ relacja bliskości między A_1 i A_n jest zachowana. Ta subtelna, ale ważna kwestia wróci do nas przy okazji omawiania własności podobieństwa, a wcześniej przy definicji skupienia.

1.3.2. Definicja. Relacja $\mathcal{R}$ w zbiorze obiektów X jest relacją równoważności, jeśli jest ona zwrotna, symetryczna i przechodnia.

Relacja równoważności w naturalny sposób grupuje obiekty do siebie podobne lub wręcz identyczne. Te wyróżnione grupy mają swoją specyficzną nazwę:

1.3.3. Definicja. Niech $\mathcal{R}$ będzie relacją równoważności w niepustym zbiorze X. **Klasą abstrakcji** elementu $a \in X$ nazywamy zbiór $(a)_R = \{x \in X : x\mathcal{R}a\}$ składający się ze wszystkich elementów zbioru X, które są w relacji $\mathcal{R}$ z a.

Odpowiednikiem klasy abstrakcji jest skupienie. Naturalność tej analogii wynika z faktu, że klasa abstrakcji skupia za pośrednictwem relacji równoważności obiekty pod pewnym względem (abstrahuje pod pozostałych cech - stąd abstrakcja). Warto podkreślić, że relacja równoważności niezwykle surowo definiuje podobieństwo między obiektami. Dzięki temu podział na jednorodne grupy jest niezwykle wyraźny. W przypadku, gdy "naturalne" grupy rzeczywiście tworzą spójne i rozłączne skupiska, kryterium to może pożądane. Jednak w większości przypadków, albo wyraźny podział nie istnieje albo empiryczna struktura nie jest znana.

Precyzja matematycznej terminologii w kontekście analizy skupień jest jednocześnie zaletą i wadą. Jej mocną stroną jest wyznaczenie niedoścignionego typu idealnego sposobu grupowania obiektów przy zachowaniu wysokiego stopień ogólności. Niestety, z drugiej strony, niedoskonały charakter naszego pomiaru (a być może obserwowanej rzeczywistości) sprawia, że powyższy model staje się wysoce niepraktyczny. Wszystko zależy od przyjętej definicji podobieństwa czy bliskości.

Wcześniej posłużyliśmy się pojęciem *wyraźny podział* nie sygnalizując jego matematycznego sensu:

1.3.4. Definicja. Niech $\mathcal{B}$ będzie rodziną podzbiorów niepustego zbioru X. Powiemy, że rodzina ta jest podziałem zbioru X, jeśli jest złożona ze zbiorów $B_i \in \mathcal{B}$ niepustych, parami rozłącznych i jej sumą jest cały zbiór X, to znaczy, jeśli spełnia następujące warunki:

- $\mathcal{B} \neq \emptyset$
- $\cup B_i = X$
- $B_i \cap B_j = \emptyset, \forall i \neq j$

Zbiory tej rodziny nazywamy **blokami podziału**. Sam podział będziemy oznaczać przez $\mathcal{B}_k$, gdzie k oznacza liczbę bloków podziału. Zatem obiekty podobne lub w tym wypadku nierozróżnialne tworzą jednorodne klasy, zwane blokami. Te zaś są rozłączne i w sumie wyczerpują cały zbiór obiektów. Wiedza na temat przynależności obiektu do danego bloku w zupełności wystarcza do opisu charakterystyk obiektu.

Naturalnym podziałem na bloki jest podział za pomocą relacji równoważności. W tym przypadku klasy abstrakcji są szczególnym przypadkiem bloków Skupiając bowiem w ramach jednego podzbioru identyczne obiekty, gwarantujemy sobie, że nie będą one należeć do żadnego innego podzbioru zawierającego również identyczne obiekty względem relacji $\mathcal{R}$.

1.3.2. Relacje w zbiorze a analiza danych relacyjnych

Elementy przedstawionego aparatu matematycznego, jeśli w ogóle występują we współczesnej analizie skupień, to jedynie w szczątkowej postaci. Mam nadzieję, że do niektórych z tych pojęć uda nam się zrobić sensowne odwołania. Metodologia zna jednak przypadki, w których algebra relacji należy do podstawowego słownika pojęciowego. Mam tu na myśli metody wykorzystujące **dane relacyjne**, tj. takie, w których relacje między obiektami są *pierwotne* wobec innych miar podobieństwa. W klasycznej metodzie pomiaru badane obiekty musimy najpierw zmierzyć, aby wyprowadzić relacje między nimi. W metodzie relacyjnej, relacje dane są a priori i to na ich podstawie możliwe jest tworzenie relacji wyższego rzędu. O ile oryginalne, obserwowalne relacje niekoniecznie muszą być relacjami równoważności, to zakładane nieobserwowalne relacji wyższego rzędu już takimi są.

Nauki społeczne wykształciła własną metodę analizy tego typu danych. Jest nią **socjometria** lub bardziej współcześnie **analiza sieci społecznych**, a podstawowym jej narzędziem są tzw. **modele blokowe**. Szczegółowe ich omówienie znacznie wykracza poza ramy tej pracy. Warto jednak o nich wspomnieć dlatego, że zrozumienie mechanizmu, który stoi za modelowaniem blokowym niewątpliwie ułatwi zrozumienie idei stojącej za analizą cechy ukrytej i skalowania. Te z kolei posłużą nam do zrewidowania postulatów stawianych pod adresem dobrej metodzie analizy skupień.

Przystępne wprowadzenie do modelowania blokowego można znaleźć w [16] lub [9]. Tutaj opowiemy o nim jedynie w zarysie.

W sieci społecznej mamy do czynienia z dwoma typami relacji: między aktorami (obserwacjami) i są to relacje niższego rzędu oraz między blokami lub pozycjami (skupieniami) i są to relacje rzędu wyższego. O pozycjach należy myśleć jako o klasach abstrakcji pewnej relacji równoważności. W tym przypadku będziemy mówić o tzw. równoważności strukturalnej.

1.3.5. Definicja. Dwie obserwacje x, y są strukturalnie równoważne wtedy i tylko wtedy, gdy mają identyczne profile relacji niższego rzędu względem pozostałych obserwacji tzn. gdy zbiory obserwacji, z którymi x i y są w relacji pokrywają się oraz gdy x i $y$następnikami relacji dla tego samego zbioru obserwacji.

Równoważnie, jeśli patrzymy na surową macierz relacji, to jednostki równoważne strukturalnie mają identyczne elementy w wierszach i kolumnach. Głównym zadaniem modelowania blokowego jest odtworzenie wyjściowej macierzy za pomocą relacji wyższego rzędu. Jak się przekonamy w dalszej części, tak postawiony problem jest podobny do analizy cechy ukrytej. Interesuje nas bowiem pytanie, czy obserwowane interakcje między aktorami można opisać za pomocą niewielkiej liczby ukrytych parametrów.

W rzeczywistości rzadko zdarza się, aby w sieci społecznej istniały obserwacje identyczne. Dlatego wygodniej jest posługiwać się pojęciem **słabej równoważności strukturalnej**. Opisuje ona stopień podobieństwa między obiektami na podstawie różnic w profilach relacji. Dla uproszczenia przyjmijmy, że relacje mają charakter binarny (0 oznacza brak relacji, a 1 jej wystąpienie). Wówczas za miarę podobieństwa między dwoma obiektami, możemy przyjąć liczbę zgodnych wystąpień zer lub jedynek na odpowiednich miejscach. W praktyce wykorzystuje się odległość euklidesową:

$$d_{ij} = \sqrt{\sum_{m=1}^{n} (x_{im} - x_{jm})^2 + (x_{mi} - x_{mj})^2}$$

Niekiedy nazywana jest ona **odległością społeczną**, ponieważ uwzględnia ewentualną asymetrię pojedynczych relacji. W ten sposób przechodzimy od dziedziny niemierzalnych relacji do odległości euklidesowych. Jeśli dwa obiekty są identyczne (równoważne strukturalnie w ścisłym sensie) to ich odległość społeczna jest równa zero. Jednak możemy przyjąć pewien zakres tolerancji α i uznać dwie jednostki za równoważne jeśli d_{ij}odległość między nimi nie przekracza poziomu α.

Kolejnym krokiem jest takie uporządkowanie obiektów, aby tworzyły one jednolite i możliwie rozłączne klasy (bloki). Jednolitość bloku można poznać po zróżnicowaniu (wariancji) występujących w nim elementów (zer lub jedynek). Po dokonaniu optymalnej permutacji, wyjściowa macierz redukowana jest do macierzy rozmiaru równego liczbie bloków. Pozostaje jeszcze zbadać, w jakim stopniu redukcja przyczyniła się do utraty wyjściowej informacji. Do tej kwestii wrócimy jeszcze w następnym rozdziale przy okazji dyskusji nad kryteriami optymalnego podziału.

1.4. Czy istnieje metoda prowadząca do *dobrego podziału*?

W pracy B.S.Everitta na temat nierozwiązanych problemów analizy skupień [22, s.177] czytamy, że rosnąca liczba metod analizy skupień doprowadziła do poważnych rozważań nad frapującą kwestią wyboru w pewnym sensie tej "najlepszej". Samo pojęcie "dobrej" metody analizy skupień może budzić pewne wątpliwości. Na pytanie, która z szerokiej gamy metod jest najlepsza Bailey [6, s.108] odpowiada, że wybór powinien zależeć przede wszystkim od intencji badacza. Utrzymuje on, że nie ma idealnej metody dostosowanej do każdego celu, co więcej każda z metod jest poprawna, jeśli jest odpowiednio zastosowana.

Brak jednoznacznych kryteriów dobrej metody analizy skupień może brzmieć zniechęcająco. Zaprezentowany poniżej sposób rozwiązania tego problemu polega na sprowadzeniu go do zagadnienia znanego i ugruntowanego w literaturze . Mianowicie, pokażemy, że analiza skupień jest **szczególnym przypadkiem skalowania** pewnego nieobserwowalnego konstruktu. Utożsamienie problemu skalowania i analizy skupień poprzedzimy krótkim opisem podstawowych idei oraz pojęć.

1.4.1. O skalowaniu cech ukrytych

Istnieje wiele społecznych konstruktów, którymi wprawieni socjologowie posługują się swobodnie bez konieczności ich definiowania. Takimi pojęciami są np. status społeczny, liberalizm, skłonność do nałogów, inteligencja czy styl życia. itp. Wiele z tych pojęć jest intuicyjnie zrozumiała dla osób nie zajmujących się zawodowo socjologią. Mniej więcej każdy z nas wie, na czym polega liberalizm, ale zapytani o jego definicję w najlepszym wypadku ograniczmy się wymieniania jego przejawów (ang. manifests). Jednak szybko okaże sie, że wybór odpowiednich obserwowalnych cech jest skuteczny, aby ustalić, w jakimś stopniu dana osoba jest liberalna.

W Encyklopedii Socjologicznej PWN, pod hasłem "skalowanie" [8] czytamy, że oprócz zbioru obserwacji do budowy modelu skalowania niezbędne są trzy elementy. Są nimi zmienne obserwowalne zwane **wskaźnikami**, zmienne nieobserwowalne zwana **cechą ukrytą** oraz typ relacji wiążące te dwa zbiory zmiennych. Ze względu na poziom pomiaru każdej ze zmiennych

można wyróżnić różne kombinacje prowadzące do różnych modeli skalowania. Klasyfikacja niektórych z nich została ujęta w formie tabeli 1.3.

Widać zatem, że model skalowania można podzielić na dwa nurty, w zależności od charakteru związku między cechą ukrytą, a wskaźnikami. Fachowym terminem jest tu **relacja wskazywania**. Do pierwszego nurtu zalicza się modele z deterministyczną relacją wskazywania, do drugiego zaś te z wersją probabilistyczną. Przykładem pierwszej grupy jest skalogram Guttmana z 1950 roku zakładający tzw. kumulatywność reakcji. Typowymi modelami z drugiej grupy są skalogramy Rascha i Mokkena.

W modelach wykorzystujących rachunek prawodpodobieństwa pojawia się pojęcie **funkcji prawdopodobieństwa reakcji** lub po prostu **funkcji reakcji** na wskaźnik. W niektórych miejscach można spotkać się z równoważnym pojęciem **krzywej charakterystycznej wskaźników** (and. indicator characteristic curve) lub **linii śladu** (trace-line). Jawna postać tej funkcji sprawia, że można ją opisać za pomocą skończonej liczby parametrów [40].

1.4.2. Podstawowe założenia modelu skalowania

Niezależnie od specyfikacji modelu struktury ukrytej istnieją pewne ogólne założenia. Szczegóły odnośnie tych założeń można znaleźć m.in w [11,27,43,44]. W tej pracy ograniczymy się jedynie kwestii fundamentalnych.

1. Cecha ukryta jest zmienną nieobserwowalną. Jej rozkład może być opisany funkcją (gęstości) prawdopodobieństwa, ale jej charakter należy do sfery hipotez.

2. Istnieje skończony zbiór obiektów, dla których estymowana jest wartość cechy ukrytej.

3. Istnieje skończony zbiór wskaźników, które są emanacją cechy ukrytej, tzn. istnieje jednoznaczna relacja między poziomem cechy ukrytej, a wartościami wskaźników.

4. Dla każdej wartości cechy ukrytej możemy wyznaczyć prawdopodobieństwo uzyskania danego profilu odpowiedzi.

5. Wskaźniki są niezależne stochastycznie względem każdego poziomu cechy ukrytej.

Punkty (1)-(2) w świetle wcześniejszego opisu nie wymagają dodatkowego komentarza. Punkty (3) i (4) dotyczą relacji między każdym ze wskaźników i cechą ukrytą. W procesie skalowania nosi ona specjalną nazwę - **relacji wskazywania**. W modelu klasy ukrytej Lazarsfelda relacja ta jest opisana przez tzw. **linię śladu** (ang. trace-line) lub **struktor** (ang. structor). W pierwotnej wersji funkcja ta miała opisywać zależność między poziomem cechy ukrytej, a prawdopodobieństwem udzielenia "poprawnej" odpowiedzi na dane pytanie. Pomysł łatwo daje się uogólnić na zależność między wartością cechy ukrytej, a daną wartością wskaźnika. Postać tej zależności nie jest znana i i jak podkreśla Lazsarsfeld [43] jest to jedno z głównych zadań analizy ukrytej struktury.

W innym miejscu Lazarsfeld [44] podaje przykłady funkcji opisywanych za pomocą jednego lub kilku parametrów. Mimo, że z ich postaci nie wynika wprost charakter relacji (probabilistyczny lub deterministyczny) to zgodnie z logiką funkcje te powinny być ciągłe i monotoniczne (im większy poziom kompetencji, tym większa szansa na dobrą odpowiedź). W wielu przypadkach przypominają dystrybuanty pewnych rozkładów.

Punkt (5) jest ostatnim, ale najważniejszym i najbardziej konkretnym założeniem modelu, gdyż rzeczywiście nakłada na niego pewne ograniczenie. Lokalną niezależność można rozumieć dwojako. Intuicyjnie, oznacza ona, że poziom cechy ukrytej dla pojedynczego respondenta nie

Tabela 1.3: Klasyfikacja metod skalowania

typ relacji	poziom pomiaru wskaźników	poziom pomiaru cechy ukrytej			
		dychotomiczny	nominalny	porządkowy	interwałowy
deterministyczny	dychotomiczny	.	analiza klas ukrytych	.	.
	nominalny	.	.	.	
	porządkowy	.	.	.	
	interwałowy	.	analiza profili ukrytych	.	
probabilistyczny	dychotomiczny	.	analiza skupień	.	
	nominalny	.		.	
	porządkowy	analiza czynnikowa danych kategorialnych		.	
	interwałowy	modele logitowe		modele probitowe	analiza czynnikowa

ulega zmianie w procesie wypełniania testu, a odpowiedzi udzielane są niezależnie od siebie. Formalnie, na łączny rozkład wskaźników narzucony jest pewne ograniczenie.

Na pierwszy rzut oka ten aksjomat może budzić pewne wątpliwości. Skoro wszystkie wskaźniki wyrażają różne miary jednego pojęcia to w pewnym sensie muszą być ze sobą powiązane (por. [61]). Jak wyjaśnia Lazarsfeld [43, s.395] statystyczna zależność między wskaźnikami determinowana jest przez ich linie śladu. Jak dodają Arminger i Kusters [5, s.376] zasada lokalnej niezależności wskaźników wynika z faktu, że cała struktura związku między nimi jest generowana (ang. generated) [3] przez wspólną cechę ukrytą. Z postulatu lokalnej niezależności płynie niezwykle ciekawa definicja **jednorodności** (ang. homogeneity) klas ukrytych. Otóż sens tej jednorodności polega na przystawaniu do pewnego schematu zależności. Jak przyznaje Gibson [27] doskonała jednorodność klas ukrytych nie jest ani możliwa, ani niezbędna, o ile *odchylenia* od *klasowej normy* będą miały charakter losowy.

1.4.3. Fundamentalne problemy skalowania

W dalszej części słownikowej definicji PWN możemy zapoznać z listą zagadnień, które dobra metoda skalowania powinna umieć rozstrzygać przekonująco tzn. w sposób statystycznie uzasadniony. Problemy te można podsumować w następujący sposób:

1. Czy zbiór wskaźników jest skalowalny tzn. czy istnieje zbiór parametrów będący w stanie odtworzyć łączny rozkład wskaźników? Jaka jest miara dopasowania modelu do danych (ang. goodness-of-fit)?

2. Ile cech ukrytych lub wymiarów zmiennej ukrytej trzeba założyć aby dany zbiór wskaźników był skalowalny?

3. Jak przyporządkować obiektom wartości zmiennej ukrytej?

4. Czy w zbiorze są wskaźniki z których bez szkody dla skalowalności można zrezygnować?

Nie wszystkie modele skalowania dostarczają przekonującej odpowiedzi na powyższe pytania. Jako przykład weźmy model analizy klas ukrytych P.F. Lazarsfelda. Zgodnie z tabelą 1.3 zakłada on nominalny charakter cechy ukrytej, która w tym przypadku nazywa się **klasą ukrytą** oraz dowolny charakter wskaźników. W niektórych miejscach można znaleźć rozróżnienie modelu na **analizę klas ukrytych** oraz **analizę profili ukrytych**, ale w tym miejscu potraktujemy te dwa modele jako równoważne.

1.5. Analiza klas ukrytych P.F. Lazarsfelda

> Głównym tematem analizy ukrytej struktury jest wnioskowanie na temat małej liczby ukrytych właściwości na podstawie dużej liczby obserwowanych wskaźników. P.F.Lazarsfeld [43, s.392]

Naturalna jest zatem obserwacja pojawiająca się u O.Wagner [61], że analiza struktur ukrytych przetrwała w rozproszeniu - w specjalnych, interesujących przypadkach, bardzo różnorodnych zastosowaniach, w licznych dziedzinach nauk społecznych. Pojęcie struktury ukrytej (ang. latent structure) jest prawdopodobnie najszerszym określeniem, ponieważ nie specyfikuje charakteru zmiennych. Arminger i Kusters ([5, s.370]) wyróżniają pięć różnych

[3] W niektórych miejscach można też spotkać termin *governed*

modeli struktury ukrytej natomiast w [61, s.11] możemy spotkać się z dwuwymiarową klasyfikacją, która zwraca cztery różne modele w zależności od poziomu pomiaru wskaźników i cechy ukrytej.

Analiza klas ukrytych jest zatem tylko szczególnym przypadkiem analizy struktur ukrytych. Za autorów tej koncepcji uważa się P.F. Lazarsfelda i N.Henry'ego którzy w 1968 napisali klasyczną pracę [44]. Warto jednak podkreślić, że już w 1959 roku ten termin pojawia się u Gibsona ([27]). Koncepcja przyciągnęła pozostałych badaczy, zwłaszcza w dziedzinie psychometrii, psychologii i badaniach edukacyjnych. Jak podaje [6] praca Lazarsfelda była również pierwszym socjologicznym spojrzeniem na problem klasyfikacji.

Model ten stanowił inspirację dla badaczy zajmujących się probabilistycznym podejściem do analizy skupień. Obecnie rozwinięcie koncepcji Lazarsfelda doczekało się kilku terminów np. **analiza skupień metodą klasy ukrytej** (ang. latent class-cluster analysis), **modelowa analiza skupień** (ang. model-based clustering), **analiza skupień oparta na modelu mieszanym** (ang.mixture-model clustering), czy **uczenie się bez nadzoru** (ang.unsupervised learning).

1.5.1. Opis modelu

Ogólne sformułowanie klasycznego modelu Lazarsfelda oraz oznaczenia, które prezentujemy poniżej jest zaczerpnięte z tekstu Gibsona [27, s.232]. Opiera się ono na następującym układzie równań.

$$\begin{cases} n & = n_1 + n_2 + \cdots + n_q \\ n_j & = n_1 p_{1j} + n_2 p_{2j} + \cdots + n_q p_{qj} \\ n_{jk} & = n_1 p_{1jk} + n_2 p_{2jk} + \cdots + n_q p_{qjk} \\ n_{jkl} & = n_1 p_{1jkl} + n_2 p_{2jkl} + \cdots + n_q p_{qjkl} \\ \cdots & \\ n_N & = n_1 p_{1N} + n_2 p_{2N} + \cdots + n_q p_{qN} \end{cases} \tag{1.1}$$

Lewa strona równości reprezentuje empiryczne liczebności różnych kombinacji wartości zmiennych dychotomicznych. W całej zbiorowości liczącej n obserwacji, n_j wskazało wartość 1 dla zmiennej o numerze j, n_{jk} obserwacji wskazało 11 dla dwóch zmiennych itd. Ostatni wiersz zawiera indeks N, który oznacza, że zostały wzięta pod uwagę całkowita liczba N zmiennych wskaźnikowych, a n_N oznacza liczebność obserwacji, które uzyskały profil złożony z samych jedynek. Warto zaznaczyć, że powyższy układ równań jest bardziej rozbudowany; w drugim wierszu po prostu wybrano jedną konkretną zmienną. Liczba równań tej postaci jest równa N ogólnej liczbie zmiennych. W ogólności, liczba równań dla obserwacji z k jest równa liczbie k-elementowych podzbiorów zbioru N-elementowego czyli $\binom{N}{k}$

Prawa strona pierwszego równania składa się z sumy liczebności q klas ukrytych. Współczynniki p_{qj} zwane **prawdopodobieństwami ukrytymi** (ang. latent probabilites) wyznaczają częstości obserwacji w klasie q, które wskazały jedynkę dla zmiennej j. Analogicznie definiuje się współczynniki z dłuższymi indeksami.

Powyższe równania opisują związek między wskaźnikami, a cechą ukrytą bez żadnych dodatkowych założeń. Warto zauważyć, że jego liniowy charakter jest immanentną cechą modelu ponieważ liczebność każdej klasy można zapisać jako wypukłą kombinację liczebności ważoną prawdopodobieństwami. Dla odróżnienia, są modele (np. w analiza czynnikowa), gdzie charakter ten jest fundamentalnym założeniem.

Aby wykorzystać powyższy model do klasyfikacji obserwacji niezbędne jest założenie (5). Zatem wystarczy, aby spełniony był wymóg niezależności wskaźników wewnątrz każdej klasy ukrytej, co prowadzi do następujących zależności:

$$\begin{cases} p_{1jk} & = p_{1j}p_{1k} \\ p_{2jk} & = p_{2j}p_{2k} \\ \dots & \\ p_{qjk} & = p_{qj}p_{qk} \\ p_{1jkl} & = p_{1j}p_{1k}p_{1l} \\ p_{2jkl} & = p_{2j}p_{2k}p_{2l} \\ \dots & \\ p_{qjkl} & = p_{qj}p_{qk}p_{ql} \\ \dots \end{cases} \tag{1.2}$$

Połączenie obydwu układów równań prowadzi do następujących zależności:

$$\begin{cases} n & = n_1 + n_2 + \dots + n_q \\ n_j & = n_1p_{1j} + n_2p_{2j} + \dots + n_qp_{qj} \\ n_{jk} & = n_1p_{1j}p_{1k} + n_2p_{2j}p_{2k} + ... + n_qp_{qj}p_{qk} \\ n_{jkl} & = n_1p_{1j}p_{1k}p_{1l} + n_2p_{2j}p_{2k}p_{2l} + \dots + n_qp_{qj}p_{qk}p_{ql} \end{cases} \tag{1.3}$$

Jeśli założymy pewną ustaloną liczebność zbioru obserwacji n, wówczas problem analizy klasy ukrytej dla s zmiennych zero-jedynkowych sprowadza się do problemu wyznaczenia $(q-1+qs) = q(1+s)-1$ parametrów, gdzie q oznacza liczę klas ukrytych oraz s oznacza liczbę warunkowych prawdopodobieństw w każdej klasie ukrytej. Z drugiej strony liczba wszystkich możliwych rozkładów łącznych wynosi $\sum_{k=1}^{s}\binom{s}{k} = 2^s$. Oznacza to, że przy ustalonej liczebności n trzeba wyznaczyć $2^s - 1$ komórek rozkładu łącznego.

Układ równań ma jednoznaczne rozwiązanie gdy liczba niewiadomych jest równa liczbie równań spełniona jest równość:

$$q(1 + s) = 2^s \tag{1.4}$$

Jak zauważa Wagner [61, s.76] *dla niewielkiej liczby klas ukrytych (2,3,4) warunek ten sprowadza się do tego, aby zmiennych jawnych było o 1 więcej niż ukrytych, a dla większej liczby klas ukrytych - w zupełności wystarczy, gdy zmiennych jawnych będzie tyle samo, co klas ukrytych.* Jest to jedynie warunek konieczny, aby móc mówić o identyfikowalności. Nie zawsze jednak dopuszczalna jest każda liczba klas ukrytych, lecz jest ona funkcją liczby wskaźników. Wyznaczenie rozwiązania wypisanego układu równań często nazywane jest metodą wyznacznikową [28]. Zapisanie go w postaci macierzowej (za pomocą tzw. macierzy bazowych, por. [61] pozwala nam określić dodatkowy warunek, który wymaga, aby zmiennych ukrytych nie było więcej niż $\frac{s+1}{2}$.

1.5.2. Problemy związane z modelem klas ukrytych

Podstawowym problemem związanym z modelem klasy ukrytej jest jego deterministyczny charakter. Powoduje to, że w wyniku procedury wyznaczania ukrytych prawdopodobieństw

nie mamy możliwości weryfikacji, jak dalece uzyskany rezultat różni się od wyjściowej struktury wskaźników tj. czy wartość *goodness-of-fit* nie różni się istotnie od 0. Innymi słowy, różnice między rozwiązaniem modelowym a empirycznym są dostrzegalne, ale na gruncie statystycznym trudno ocenić ich wielkość.

Gdybyśmy bowiem chcieli wykorzystać do tego celu statystykę X^2 Pearsona, okazałoby się, że liczba stopni swobody jest równa 0, ponieważ liczba parametrów w „modelu nasyconym" jest równa liczbie parametrów modelu z klasami ukrytymi.

W przypadku, gdy zakładana liczba parametrów jest mniejsza niż 2^s mamy do czynienia z sytuacją modelowania, wobec czego konieczne jest wykorzystanie procedur estymacyjnych. W deterministycznym modelu Lazarsfelda nie jednak żadnych wskazówek na temat kształtu estymatorów parametrów. W wielu przypadkach może powodować to niejednoznaczność przyporządkowań obiektów do klas ukrytych.

Załóżmy jednak, że bylibyśmy w stanie obliczyć warunkowe prawdopodobieństwo przynależności obiekty do danej klasy. Pojawia się wówczas dodatkowy problem, w jaki sposób przełożyć uzyskany wynik na ostateczne przyporządkowanie.

1.6. Postulaty dobrej metody analizy skupień

Będziemy teraz chcieli przetłumaczyć powyższe problemy na grunt analizy skupień, mając na uwadze specyfikę tej ostatniej metody. Jej wyjątkowość polega przede wszystkim na szczególnym przypadku cechy ukrytej. Jest nią wartość klasyfikacji. Zmienna ta jest jednowymiarowa i mierzona jest na skali nominalnej. Wskaźniki, które służą do jej wyznaczania mogą być mierzone na dowolnej skali, co więcej nie jest wymagana ich jednorodność (tj. część może być interwałowa, część porządkowa lub nominalna). W tej pracy ograniczymy się jednak do wskaźników jednorodnych, gdyż już na tym etapie pojawiają się pewne trudności związane z modelowanie. Warto jednak pamiętać, że możliwy jest również alternatywny model z mieszanymi wskaźnikami.

Za pomocą komentarza do powyższych postulatów chcieliśmy pokazać, że między problemem skalowania, a problemem analizy skupień występuje istotna analogia. W dalszej części pokażemy, że faktycznie mamy do czynienia z czymś więcej. Jak zobaczymy, **analiza skupień jest szczególnym przypadkiem skalowania cechy ukrytej** na podstawie obserwowalnych wskaźników. W ten sposób postulaty formułowane pod jej adresem w naturalny sposób powinny odzwierciedlać postulaty stawiane metodom skalowania. Dlatego o danej metodzie analizy skupień powiemy, że jest dobra, w momencie gdy dostarczy statystycznie uzasadnionej odpowiedzi na każde z poniższych pytań:

1. Czy zbiór obiektów jest **segmentowalny**? Jak dobrze model pozwala odtwarzać strukturę zbioru obiektów?
2. Ile skupień należy założyć aby uzyskać odpowiedni stopień dopasowania modelu do danych?
3. Jak przypisać obiekty do odpowiednich skupień?
4. Czy w zbiorze istnieją obiekty, które można wyeliminować?

Ad.1 O ile w zagadnieniu skalowania szukamy relacji między wskaźnikami, a cechą ukrytą o tyle w analizie skupień, chcemy przypisać obserwacje do odpowiednich klas. Gdyby jednak potraktować zbiór obserwacji jako zbiór profili - czyli rozkładów łącznych wskaźników, wówczas nasze zagadnienie nie różni się od problemu skalowania. Odtwarzanie łącznego rozkładu

wskaźników można sprowadzić do pytania, na ile dwie informacje: na temat przynależności obiektu do danego skupienia oraz na temat odległości między skupieniami pozwala nam odtworzyć wyjściowe relacje podobieństwa (odległości) między dowolną parą obiektów.

W tym miejscu uzasadnione jest skojarzenie z dwoma znanymi modelami. Pierwszym z nich jest model czynnikowy, w którym na podstawie korelacji między czynnikami staramy odtworzyć się strukturę korelacji między wskaźnikami. Drugi to modele blokowe, w którym szukamy takiego wzoru strukturalnego, dzięki któremu zrekonstruujemy relacje niższego rzędu między dowolną parą obiektów.

Do mierzenia jakości odtwarzania najczęściej wykorzystywana jest pewna funkcja straty lub błędu (ang. loss lub error function). Jej szczególnym przypadkiem jest statystyka X^2 Pearsona. Jednak w wielu sytuacjach, gdy analiza skupień nie jest traktowana jako model skalowania wprowadza się inne rodzaje funkcji straty. W ogólnej postaci posiadają one dwa argumenty, które są ze sobą porównywane: empiryczne i modelowe rozkłady liczebności. W zależności od tego, w jaki sposób zdefiniowana będzie funkcja błędu (np. według modalnej, modułowa, kwadratowa) różne będą kryteria optymalności podziału.

Ad.2 Większość modeli z cechą ukrytą o których wspomnieliśmy miało charakter jedno lub dwuwymiarowy. Model analizy skupień zawsze jest jednowymiarowy w sensie liczby klas ukrytych. Dlatego pytanie o liczbę klas ukrytych w analizie skupień może wydawać się źle postawione. Bardziej sensownym wydaje się pytanie o jej wymiarowość. Jeśli wymiar klasy ukrytej wynosi 1, oznacza to, że zbiór nie jest segmentowalny (równoważnie, jest amorficzny lub nie posiada żadnej struktury). Jak pokażemy w następnych rozdziałach, wymiar przestrzeni ukrytej może być rozwiązany za pomocą metod statystyki inferencyjnej.

Ad.3 Tym, co nas najbardziej interesuje w procesie skalowania jest indywidualny poziom cechy ukrytej dla każdej obserwacji. Surowy model matematyczny definiuje sposób przyporządkowania elementu do bloku podziału bez kształtu tej funkcji. W niektórych przypadkach znamy kształt tego związku. Tak jest np. w skalogramie Rascha (krzywa logistyczna) czy modelu czynnikowym (liniowość). W przypadku analizy skupień na ogół ten kształt nie jest znany. Dlatego bez dodatkowych założeń na temat modelu problem analizy skupień może być nierozwiązywalny na gruncie statystycznym, jeśli nie jest dobrze zdefiniowany.

Ad.4 W każdym zestawie pytań zdarzają się pytania, które są niemiarodajne (np. jeśli mowa o teście, to są to pytania ekstremalnie łatwe lub trudne). O tym, czy dane pytanie wykazuje taką własność dowiadujemy się na przykład badając jego brzegowy rozkład. Mała wartość wariancji wskaźnika, świadczy o tym, że ma on niską moc dyskryminacyjną, co oznacza, że uzyskalibyśmy podobną jakość modelu bez użycia tego wskaźnika. Obecność takich pytań jednak może zaburzać średni wynik testu.

W analizie skupień natomiast często mamy do czynienia z obserwacjami, które wyraźnie różnią się od reszty zbioru. Są one nazywane jednostkami odstającymi (ang. outliers). Obecność takich obserwacji również może zakłócać identyfikację prawdziwej struktury zbioru. Rozpoznanie takich jednostek nie zawsze jest do zrealizownia za pomocą standardowych metod. Jak pokażemy w następnych rozdziale, każda decyzja odnośnie manipulacji takimi obserwacjami powinna być dobrze uzasadniana.

Wiele z tych problemów może choć nie musi przekształcić się w statystyczne problemy estymacji lub testowania hipotez. Jednak, jak pokażemy w dwóch następnych rozdziałach, założenie probabilistycznego związku obserwowalnych reakcji ze zmienną ukrytą oferuje wiele praktycznych narzędzi do dokonania obiektywnej oceny. Niemniej jednak, nie oznacza to, że

zawsze są to problemy łatwo rozstrzygalne na gruncie statystyki, co pokażemy w rozdziale trzecim i czwartym. Jeśli jednak te kwestie są rozstrzygalne, wówczas dostarczają dobrze uzasadnionych odpowiedzi.

Rozdział 2

Podstawowe problemy analizy skupień

2.1. Specyficzne problemy analizy skupień

2.1.1. Definicje skupienia

> Powinna istnieć jasny, wyraźny oraz intuicyjny opis klasyfikacji, skupienie powinno coś oznaczać. Niektóre opublikowane metody analizy skupień mają ładne algorytmy, ale gdy już zakończą swój przebieg, trudno jest zobaczyć, jaki właściwie problem został rozwiązany. [37, s.242]

Definicja skupienia jest jednym z poważniejszych problemów, z którymi musi zmierzyć się jej użytkownik. Problem jest dwojakiego rodzaju: z jednej strony nie ma jasnych wskazówek dla określenia granic skupień, z drugiej zaś brakuje uzasadnień dla decyzji, które obserwacje powinny być włączone w ramach odpowiednich skupień. Nie istnieją mocno ugruntowane reguły dla zdefiniowania skupienia. Aktualnie używane różnią się ze względu na dyscyplinę lub cel, jaki przyświeca badaczowi.

Ponieważ samo skupienie nie jest dobrze zdefiniowanym pojęciem, nie można przedstawić żadnych formalnych zasad odnośnie znajdywania skupień. Przeglądając literaturę na temat analizy skupień, nie sposób się nie zgodzić z tym sformułowaniem. Ling [45, s.159] twierdzi, że w wielu dostępnych metodach segmentacji, skupienie nie jest zdefiniowane *explicite*. Jak zauważa [24, s.6], jeśli chodzi o pojęcie skupienia, grupy czy klasy są one zwykle używane w sposób zupełnie intuicyjny i bez żadnej próby formalnej definicji. Z drugiej strony podkreśla, że w rzeczywistości formalna definicja nie tylko jest trudna do wyrażenia, ale również łatwo o jej zbyt wąski charakter w odniesieniu do tak szerokiego zakresu problemów. Przed zbytnią formalizacją przestrzega również Cormack [18, s.329]: *Nieostre definicje skupienia będą dopuszczać wiele spośród wielowymiarowych bytów (ang. multidimensional amebhoas) (...) Bez żadnej formalnej definicji wszystko może być dyskusyjne i dopuszczalne.* Jest zdania, *że zasadniczym celem analizy skupień powinno być opisanie danych w sposób prostszy niż występują one w oryginale bez stosowania przesadnego aparatu matematycznego.*

Czy przejście do porządku dziennego nad bezradnością formalizacji definicji skupienia może mieć jakieś niekorzystne konsekwencje?

Według Klastorina [39, s.92] problem klasyfikacji jest po prostu *źle postawiony* (ang. ill-defined). Zdaniem autora, brak definicji skupienia doprowadził do niekontrolowanego rozwoju procedur analizy skupień, z których większość posiada w bardzo małym stopniu rozwinięte podstawy teoretyczne. Niemniej krytyczna jest uwaga Punja [52, s.134] odnośnie faktu, że

wskaźnikiem ogólnego braku zrozumienia metodologii klasyfikacji jest błędne określanie przez wielu badaczy, jaka **metoda** została przez nich użyta. Do mało optymistycznych wniosków doszedł również Everitt [24], twierdząc, że prawdopodobnie problem bezradności w poszukiwaniu formalnych rozwiązań, polega na tym, że faktycznie, analiza skupień nie wypracowała jak dotąd powszechnie akceptowalnej definicji skupienia. Z punktu widzenia oceny "dobrej" metody, według [6, s.112] trudno jest dokonać ewaluacji różnych metod ponieważ często opierają się o różne definicje skupienia i wybór metody przez badacza powinien opierać się o to, w jaki sposób definiuje on skupienie..

Problem ten próbowano rozwiązywać na wiele sposobów. W efekcie pojawiły się definicje o różnym stopniu ogólności i scisłości. Ponieważ znaczna część autorów nie poświęca wiele miejsca pojęciu skupienia, warto przedstawić odważny próby stworzenia precyzyjnej definicji. Poniżej ograniczymy się do podania dwóch pouczających przykładów.

Pierwszy z nich, oparty na podstawowych pojęciach z topologii został wprowadzony przez Linga w [45]. Jak podaje autor, jego koncepcja jest uogólnieniem metody pojedynczego wiązania w metodach hierarchicznych. Jedynym założeniem, które wykorzystuje do budowy ścisłej definicji skupienia jest istnienie macierzy odległości $D = d[ij]$ między obiektami. Dzięki obecności metryki możliwe jest określanie przestrzeni metrycznej (X, d). Ling zaczyna od zdefiniowania **r-łańcucha**.

2.1.1. Definicja. Ciąg elementów $x_i \in X$nazywamy **r-łańcuchem** jeśli odległość między kolejnymi elementami jest nie większa niż r.

2.1.2. Definicja. O podzbiorze C można powiedzieć, że jest **r-połączony** (ang. connected) jeśli istnieje permutacja jego elementów tworząca r-łańcuch.

2.1.3. Definicja. Zbiór C nazywany jest **(k,r)-związanym** (ang. bounded) jeśli spełnia warunek, że dla każdego elementu $x \in C$ można wskazać k-elementowy podzbiór T, że odległość między x i każdym elementem z T jest nie większa od r

2.1.4. Definicja. Jeśli C jest jednocześnie (k,r)-związany oraz r-połączony, wówczas mówimy, że jest on **k,r**-połączony.

Pojęciem, do którego dążymy na samym końcu jest skupienie. Pozostaje jeszcze jedna pomocnicza definicja:

2.1.5. Definicja. Mówimy, że podzbiór C jest **(k,r)-skupieniem**, jeśli r jest najmniejszą liczbą taką, że C jest (k,r)-połączony, a jednocześnie jest on maksymalny w sensie zawierania.

Ostatecznie, mamy, że:

2.1.6. Definicja. C jest k-skupieniem jeśli jest k, r-skupieniem dla pewnego r

Mimo, że ciąg powyższych definicji na pierwszy rzut oka może wyglądać przytłaczająco lub sprawiać wrażenie przerostu formy nad treścią, to w wielu praktycznych sytuacjach może okazać się pomocny. Definicja ta może być szczególnie przydatna dla określania metod opartych na macierzy odległości między obserwacjami (zob. następna sekcja).

Zauważmy, że parametr r może być interpretowany jako **moment powstania skupienia**. Analogicznie możemy zdefiniować moment jego rozpadu. Ling proponuję nazwę **indeks izolacji** (ang. isolation index) jako różnicę $s - r$ gdzie s oznacza parametr dla najmniejszego k-skupienia, który w sposób właściwy zawiera C. Za pomocą dwóch momentów: powstania i rozpadu (k,r)-skupienia dla dowolnych parametrów k i r można opisać strukturę skupienia, jednak, jak zauważa Ling potrzebny jest do tego specjalny algorytm. Do tej kwestii wrócimy jeszcze w następnym rozdziale, gdy będziemy mówić o **segmentowalności** zbioru.

Drugi, zupełnie odmienny pomysł definicji skupienia, opiera się na modelu probabilistycznym. Mimo, że ogólny pomysł był od dawna eksploatowany to w kontekście definicji pojawia

się po raz pierwszy w tekście [35, s.269]. Warto przyswoić sobie poniższy tok rozumowania, ponieważ pojawi się on w Rozdziale 3 w bardziej rozbudowanej wersji. Punktem wyjścia do sformułowania definicji nie jest ani macierz odległości między obserwacjami, ani też sam zbiór obserwacji. Podstawowym pojęciem jest **profil**, czyli wektor wartości zmiennych dla danej obserwacji. Każdy profil jest realizacją pewnego procesu losowego, które determinowane jest przez dane skupienie. W przeciwieństwie do koncepcji Linga, pewne skupienia istnieją zawsze, tylko nie są one obserwowalne. Skupienie jest strukturą wcześniejszą niż cały proces analityczny, który u Linga prowadził do identyfikacji skupienia. Jest ono definiowane jest jako generator lub źródło profili. W sensie statystycznym jest to po prostu zestaw parametrów, które wyznaczają wartość profilu. Innymi słowy, profil jest realizacją pewnego procesu losowego, który jest kombinacją procesów realizujących się w każdym ze skupień z różnymi częstościami.

2.1.2. Podobieństwo między obiektami

> Nie jest do końca jasne, w jaki sposób rozpoznaje się skupienie na płaszczyźnie, jednak prawdopodobnie proces poznawczy polega na ocenie względnych odległości między punktami. [22, s.13]

> W matematyce często bierze się metaforę i przekształca się ją w narzędzie matematyczne. Jakieś obrazowe zdanei z życia codziennego bierze się dosłownie i wykazuje się, że jest ono ścisłe i logiczne. Tak stało się z metaforycznym sesnem słowa odległość. [54, s.260]

Podobieństwo między obserwacjami

W [32, s.459] możemy przeczytać, że specyfikacja odpowiedniej miary podobieństwa jest o wiele bardziej istotna w uzyskaniu sensownego rozwiązanie niż wybór algorytmu.

Jeśli analiza skupień ma na celu wyodrębnienie podzbiorów złożonych z jednostek maksymalnie do siebie podobnych, należy zdefiniować, co rozumiemy poprzez owe podobieństwo (ang. similarity). Jak zauważa [35, s.271] skoro podobieństwo między dwoma profilami wygenerowanymi z tej samej przestrzeni jest kluczowe dla definicji skupienia, jest ono również kluczowe dla większości metod analizy skupień. Wagę problemu dostrzega również [30, s.359]. Jednym z głównym problemów w stosowaniu analizy skupień jest wybór miary bliskości (proximity measure) między parami obiektów. W dalszej części swojej pracy pokazuje, że na tej samej macierzy danych surowych wyniki analizy skupień mogą wyraźnie różnić się w zależności od wybranej miary podobieństwa.

Everitt [24] wyróżnia dwa możliwe sposoby pomiaru podobieństwa: **bezpośrednie i pośrednie**. Pierwszy polega na zadaniu pytań grupie ekspertów, czy dane dwa obiekty są podobne. Załóżmy, że mamy pewien zbiór obiektów, które należy ze sobą porównać parami posługując się skalą od 0 do 100, gdzie 0 oznacza zupełny brak podobieństwa, a 100 identyczność. Respondent ma za zadanie wypełnić macierz podobieństwa między obiektami podając liczbę z przedziału 0 do 100. Następnie uśrednia się wynik dla wszystkich respondentów. Podstawową wadą tej metody jest brak gwarancji, że relacje między obiektami będą przechodnie. Oznacza to, że gdybyśmy chcieli potraktować te obiekty jako punkty w przestrzeni, to uzyskana macierz nie będzie macierzą odległości euklidesowych. Jest tak, ponieważ nie spełnia ona nierówności trójkąta, a jedynie warunki minimum: nieujemność, symetryczność oraz zeruje się dla obiektów identycznych.

Dlatego w praktyce częściej wykorzystuje się metody pośrednie, polegające na transformacji surowej macierzy danych z obserwacjami w wierszach i zmiennymi w kolumnach. Jak

pisze Gordon [29, s.120], wiele metod klasyfikacji wymaga aby surowa macierz danych została przekształcona w symetryczną macierz nie-podobieństw (ang. dissimilarities). To ona stanowi podstawę do analizy skupień, podobnie jak macierz korelacji jest bazą analizy czynnikowej.

Dystans lub **odległość** jest tylko szczególnym przypadkiem podobieństwa (lub równoważnie niepodobieństwa) między obiektami, który milcząca zakłada, że nasze obserwacje możemy traktować jako pewne fizyczne obiekty w wielowymiarowej przestrzeni oraz, że istnieje jakiś punkt odniesienia (np. początek układu współrzędnych) względem którego określamy położenie naszych obserwacji.

Zwykle jesteśmy przyzwyczajeni do używania metryki euklidesowej, tzn. takiej, że odległość między dwoma punktami wyraża się wzorem:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{2.1}$$

Jest ona szczególnym przypadkiem tzw. odległości Minkowskiego z parametrem $m = 2$.

$$d(x,y) = \left(\sum_{i=1}^{m}(x_i - y_i)^p\right)^{\frac{1}{p}} \tag{2.2}$$

Z kolei dla $m = 1$ otrzymamy tzw. metrykę miejską (ang. city-block metric).

$$d(x,y) = \sum_{i=1}^{m}(x_i - y_i) \tag{2.3}$$

Decydując się na wybór metryki musimy mieć na uwadze jej podstawowe własności. Przykładowo, powyższe metryki mają tę własność, że traktują wszystkie wymiary równorzędnie. Ponadto nie uwzględniają żadnych informacji na temat zróżnicowania danej cechy. W ich definicje wpisane jest *implicite* założenie o niezależności zmiennych, co oznacza posługiwanie się ortogonalnym układem współrzędnych.

Wartą zapamiętania jest wskazówka jaką podaje Kaufmann [38, s.469]: Decydując się na dany sposób przekształcenia oryginalnej macierzy danych w pewną miarę podobieństwa między dwoma obiektami należy dokonać istotnego rozróżnienia między podobieństwem opartym na jednej tylko cesze i podobieństwem opartym na wielu cechach. Wyobraźmy sobie, że dokonujemy pomiaru obserwacji ze względu na zmienne X i Y. Obie są mierzone w tych samych jednostkach, przy czym zmienna X ma o wiele większą wariancję niż zmienna Y. Jest intuicyjnie zrozumiałym, że *nominalnie* ta sama odległość euklidesowa po zrzutowaniu na oś X będzie miała *mniejsze znaczenie* niż po zrzutowaniu na oś Y.

W [35] możemy spotkać się z ciekawą koncepcją kategorii **metryk kontekstowych** (ang. context metrics). Uwzględniają one efekt pozostałych obiektów w otoczeniu pary punktów, dla których odległość jest wyznaczana. Ten rodzaj odległości uświadamia nam fakt, że w zasadzie o dystansie powinniśmy myśleć, jako o czymś relatywnym. Prostym przykładem metryki kontekstowej jest odległość do (wybranych) sąsiadów:

Niepożądaną własnością metryki Minkowskiego jest "faworyzowanie" wymiarów, na których mierzone są zmienne o skali wyraźnie większej od pozostałych zmiennych. W obu przypadkach - różnicy skali i zróżnicowania - narzucającym się rozwiązaniem jest szeroko pojęta **standaryzacja** obejmująca również takie zabiegi jak **normalizację** czy **ważenie** wskaźników.

Innym niebezpieczeństwem jest liniowe skorelowanie wskaźników. Jeśli mierzymy odległość między obiektami za pomocą trzech zmiennych, między którymi występuje silna pozytywna korelacja, wówczas uzyskana różnica będzie zwielokrotniona w efekcie nakładania się tego samego czynnika. Efekt ten może być usunięty na różne sposoby, z których najczęściej zalecanym jest użycie **odległości Mahalanobisa**. Wyraża się ona następująco:

$$d_M(x,y) = \sqrt{(\mathbf{x}-\mathbf{y})\Sigma^{-1}(\mathbf{x}-\mathbf{y})^T} \tag{2.4}$$

W tym przypadku x i y oznaczają p-wymiarowe wektory obserwacji, a Σ jest macierzą kowariancji między p zmiennymi. Jak łatwo zauważyć, podobny zapis występuje również w wykładniku gęstości wielowymiarowego rozkładu normalnego (zamiast wektora y jest wektor wartości oczekiwanych):

$$f(x_1, x_2, ..., x_n) = (2\pi)^{-\frac{n}{2}} \cdot (\det \Sigma)^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)^T}$$

Jest tak, ponieważ stosując tę odległość zakładamy *implicite*, że rozkłady prawdopodobieństwa w każdej klasie mają postać zbliżoną do wielowymiarowych krzywych Gaussowskich.

Okazuje się, że założenie na temat rozkładu może być kluczowe dla wyznaczenia odpowiedniej metryki. Jak podają Yu i in. [64, s.533] jeśli znana jest funkcja, która generuje uzyskany łączny rozkład wskaźników, możliwe jest wskazanie metryki, która najlepiej do niego pasuje. Zgodnie z tym, co pisze [18, s.324] istnieją sytuacje, gdy ten rozkład można zidentyfikować poprzez podanie wartości oczekiwanych i kowariancji. Nawet jeśli tego nie sygnalizuje się explicite, to najczęściej zakłada się, że dany rozkład jest normalny lub wykładniczy. Jak twierdzą Yu i współautorzy [64, s.536] warto być świadomym, że są rozkłady, dla których istnieją metryki inne niż najbardziej intuicyjna, czyli euklidesowa. Ci sami autorzy konkludują, że np. dla rozkładu normalnego właściwa jest metryka euklidesowa, a dla wykładniczego miejska.

Sprowadzenie relacji podobieństwa do relacji odległości zwykle nie jest zupełnie oczywiste. Z uwagi na różnorodność skal pomiarowych, jakimi dysponujemy, musimy zadać sobie pytanie, w jakim stopniu sformułowanie podobieństwa w kategoriach macierzy odległości zachowuje "rzeczywiste" relacje między obiektami.

Na pierwszy rzut oka wydaje się, że obliczanie odległości między zmiennymi jakościowymi (mierzonymi na skalach słabszych niż przedziałowa) jest niemożliwe, ponieważ relacja podobieństwa między takimi obiektami ma charakter binarny (dwa obiekty są albo nie są podobne). A jednak literatura zna szeroką gamę metryk dla tego typu zmiennych. Opis lub porównanie wszystkich znanych i dostępnych w literaturze metod zajęłoby tu zbyt wiele miejsca. Bogata baza wiedzy na ten temat znajduje się na stronie SimMetrics [1]

Green [30] wyróżnia dwa typy przekształceń: **zachowujące informacje** (ang. information preservation measures) oraz **redukujące informacje** (ang. information reduction measures).

Do miar redukujących informacje należy druga obok odległości miara podobieństwa - odległość korelacyjna. Oparta jest ona na współczynniku korelacji ρ^2 obliczanym nie zmiennych, lecz dla obserwacji. To tak, jakby dokonać transpozycji surowej macierzy danych - zmienne stają się obserwacjami, a obserwacje zmiennymi. Dane dwa profile są do siebie podobne jeśli współczynnik korelacji liniowej jest wysoki.

Pod adresem tej miary podobieństwa można sformułować kilka zarzutów. Przede wszystkim, zmienne statystyczne i obserwacje to dwa różne obiekty, do których powinny być stosowane różne miary. Mówiąc w dużym uproszczeniu, współczynnik korelacji liniowej zdaje

[1] http://www.dcs.shef.ac.uk/ sam/simmetrics.html

sprawę, w jakim stopniu jedna zmienna jest liniową funkcją drugiej zmiennej. Trudno jednak określić, co może oznaczać korelacja między obiektami w przestrzeni. Łatwo bowiem można wskazać takie pary obiektów, dla których osiągany współczynnik korelacji jest mylący.

Tabela 2.1: Przykład obserwacji „skorelowanych liniowo", $\rho^2 = 1$, $d_{1,2} = 6$

	x1	x2	x3	x4
ω_1	1	1	1	1
ω_2	4	4	4	4

Tabela 2.2: Przykład obserwacji „liniowo niezależnych", $\rho^2 = 0$, $d_{1,2} = \sqrt{2}$

	x1	x2	x3	x4
ω_1	1	1	0	0
ω_2	1	0	1	0

Podobieństwo między skupieniami

Podobieństwo w skali mikro (między pojedynczymi obserwacjami) różni się koncepcyjnie od podobieństwa w skali makro (między agregatami obserwacji, czyli skupieniami). Jak pamiętamy z teorii modelowania blokowego, relacje na poziomie pojedynczych obiektów miały inne właściwości niż na poziomie zagregowanych bloków.

Jak podaje Bailey [6, s.69] w analizie skupień, podobnie jak w każdej technice operującej na agregatach a nie tylko pojedynczych obserwacjach, istnieje ryzyko popełnienia **błędu ekologicznego** (ang. ecological fallacy). Istotne jest pytanie, na ile z podobieństwa między skupieniami można wnioskować na temat podobieństwa między obiektami, w których się one znajdują.

Z punktu widzenia teorii zbiorów dane dwa wyróżnione podzbiory można uznać za tym bardziej odległe, im w ich części wspólnej znajduje się mniej elementów. Zauważmy, że taka definicja nie wymaga określenia żadnej metryki, dlatego też traktuje dowolne dwa rozłączne skupienia jako równo odległe nawet jeśli w fizycznym sensie różnica odległości jest znaczna.

Do mówienia o różnicach między skupieniami niezbędne jest założenie, że zbiór jest w jakimś stopniu segmentowalny tzn. możemy wyróżnić w nim co najmniej dwa skupienia. Kwestię segmentowalności zbioru dokładniej omówimy w następnym rozdziale, dlatego tutaj tylko zasygnalizujemy obecność niektórych terminów i zagadnień.

Na podobieństwo między skupieniami możemy patrzeć z dwóch perspektyw: zewnętrznej i wewnętrznej. Pierwsza z nich dostarcza sumarycznej informacji na temat odległości między segmentami. Zalicza się do nich m.in. indeks separowalności czy indeks sylwetki (zob. następne sekcje). Z kolei spojrzenie od wewnątrz pozwala ocenić wzajemne położenie dowolnej pary skupień. Pojawia się wtedy jednak istotne pytanie, w jaki sposób taką odległość mierzyć oraz co *de facto* jest reprezentacją skupienia.

Problem ten został w dużej części rozwiązany poprzez statystyków zajmujących się metodami hierarchicznymi. Wprowadzili oni różne rodzaje odległości. Do najbardziej popularnych zaliczają się: najbliższego sąsiada, najdalszego sąsiada, średniego wiązania, środków ciężkości i in. W Aneksie, w którym zawarty jest również formalny opis procedury, możemy zobaczyć, że wybór odległości między skupieniami ma nieraz duże znaczenie na przebieg procesu segmentacji.

2.1.3. Efektywność algorytmów

Rozmiar zadania

Z matematycznego punktu widzenia zadanie analizy skupień jest dość prymitywnym zagadnieniem optymalizacyjnym. Zakładamy, że każda obserwacja może należeć do jednego i tylko do jednego skupienia. Skończony zbiór obiektów implikuje skończoną liczbę możliwych podziałów, zatem przy odpowiednio zdefiniowanej funkcji celu (przykłady takich funkcji zaprezentujemy w następnym rozdziale), optymalne rozwiązanie jest zawsze osiągalne w skończonej liczbie kroków.

Złożoność zadania przewyższa jednak możliwości dzisiejszych komputerów. Aby zdać sobie sprawę z rozmiaru zadania, wystarczy podać wzory na liczbę wszystkich możliwych podziałów zbioru n - elementowego. Jak wiadomo, wyrażają się one n-tymi liczbami Bella. Do ich opisu, zwykle stosuje się wzór rekurencyjny z warunkiem początkowym $B_0 = B_1 = 1$:

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k \tag{2.5}$$

Wyrazy tego ciągu rosną bardzo szybko.[2] Już dla zbioru o 100 obserwacjach, liczba możliwych grupowań wynosi $4,7510^{115}$.

Zazwyczaj jednak możemy wykluczyć pewne rozwiązania, gdyż mogą nas interesować tylko podziały na określoną liczbę skupień. Wówczas wyznaczamy liczbę podziałów zbioru n-elementowego na k podzbiorów. Wyrażają się one przez tzw. liczby Stirlinga II rodzaju, dla których również używa się zapisu rekurencyjnego:

$$\begin{aligned} S(0,0) &= 1 \\ S(1,0) &= 0 \\ S(1,1) &= 1 \\ S(n,k) &= kS(n-1,k) + S(n-1,k-1) \end{aligned}$$

Wzór jawny, dla parametrów n i k można znaleźć np. w [24, 32, 41]:

$$S(n,k) = \frac{1}{k!} \sum_{i=1}^{k} (-1)^{k-i} \binom{k}{i} i^n \tag{2.6}$$

Dla przykładu $S(5,2) = 15, S(10,3) = 9330$, ale dla $S(100,5)$ czyli możliwych podziałów zbioru 100 elementowego na 5 skupień jest już $6,610^{67}$. Mimo, że jest to prawie dwukrotnie mniejsza liczba niż B_n to nadal wraz ze wzrostem liczebności zbioru obserwujemy **eksplozję kombinatoryczną**.

Jedynym sensownym sposobem pomiaru efektywności algorytmów jest porównanie wyników uzyskanych za jego pomocą z pewnym stanem faktycznym. W analizie skupień stan faktyczny nie jest znany i objawia nam się dopiero po zadziałaniu algorytmu. Załóżmy jednak, że struktura zbioru jest nam znana i dzięki temu jesteśmy w stanie ocenić zbieżność rezultatu naszej procedury z rzeczywistością. W tego typu badaniach symulacyjnych najprostszym, a zaraz najbardziej powszechnym jest kryterium zewnętrznym jest **indeks Randa** (ang. Rand Index).

[2] Kolejne liczby Bella dla indeksów od 1 do 500 można znaleźć między innymi na: http://www.research.att.com/ njas/sequences/b000110.txt.

Indeks Randa Idea indeksu opiera się na tzw. korespondencji (zob. [34]) między podziałami. Załóżmy, że mamy pewien zbiór obiektów X o liczebności n. Bez straty ogólności możemy założyć, że porównujemy tylko dwa podziały $U = (u_1, u_2, ..., u_p)$ i $V = (v_1, v_2, ..., v_r)$. Małe litery indeksowane kolejnymi liczbami naturalnymi oznaczają skupienia danego podziału. Zgodnie z klasyczną definicją podziału, w obu przypadkach są one parami rozłączne i wyczerpują cały zbiór. Zauważmy, że U i V są po prostu dodatkowy zmiennymi w zbiorze obiektów. Na ich podstawie możemy wyróżnić cztery grupy, pary należące do:

1. tych samych skupień zarówno w U i V
2. różnych skupień w U ale do tych samych w V
3. tych samych skupień w U ale różnych w V.
4. do różnych skupień w U i V

Obserwacje typu (1) i (2) to tzw. zgodności (ang.agreements) w klasyfikacji a obserwacje typu (3) i (4) niezgodności (ang. disagreements). Narzucająca się analogia do korelacji rangowej jest tu całkowicie uzasadniona. Tu obserwujemy jej szczególny, binarny przypadek. Wartość indeksu obliczana jest następująco:

$$RI = \frac{n_1 + n_4}{n_1 + n_2 + n_3 + n_4} \tag{2.7}$$

Indeks przyjmuje wartości z zakresu $(0, 1)$. Duże wartości świadczą o podobieństwie między dwoma uzyskanymi podziałami. Odmianą indeksu jest **indeks Jaccarda**, który bierze pod uwagę tylko zgodne pary obserwacji:

$$JI = \frac{n_1}{n_1 + n_2 + n_3} \tag{2.8}$$

Jak zauważają (Ruzzo) oba powyższe indeksy mają istotną wadę - w przypadku porównywania dwóch losowych rozkładów nie mają stałej wartości oczekiwanej. Dlatego w niektórych pakietach statystycznych (np. w R) można spotkać poprawiony indeks Randa (adjusted Rand Index) (zob. [34]), który zakłada hipergeometryczny model losowości tzn. podziały losowane są tak, aby zapewnić tę samą liczbę skupień oraz tę samą liczbę obserwacji w odpowiednich skupieniach. Autorzy pokazują, że wartość oczekiwana indeksu jest równa:

$$E\left(\sum_{i,j}\binom{n_{ij}}{2}\right) = \frac{\sum_i \binom{n_{i.}}{2}}{\sum_j \binom{n_{.j}}{2}}$$

Symbol $n_{i.}$ oznacza liczbę obserwacji należących do skupienia i w pierwszym podziale, natomiast $n_{.j}$ oznacza liczbę obserwacji należących do skupienia j w drugim podziale.

Usprawnienie indeksu polega na normalizacji względem maksymalnej i oczekiwanej wartości indeksu. Konkretnie (por. [**?**]):

$$RI_{popr} = \frac{RI - E(RI)}{\max(RI) - E(RI)}$$

Pozostałe elementy równania wyrażają się następująco:

$$RI = \sum_{i,j} \binom{n_{ij}}{2}$$

$$\max(RI) = \frac{\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}}{2}$$

Aby porównać wartości powyższych indeksów dla tego samego zbioru, zilustrujmy to prostym przykładem. Poniżej przedstawiono rozkład łączny przynależności do 3 skupień dla dwóch różnych hipotetycznych podziałów.

Tabela 2.3: Łączny rozkład liczebności dla dwóch podziałów U i V

	v_1	v_2	v_3	
u_1	3	1	0	4
u_2	0	1	1	2
u_3	0	1	3	4
	3	3	4	10

Licząc indeks Randa musimy wyznaczyć cztery wartości. Najpierw obliczmy n_1 liczbę par, które należą do tego samego skupienia w obu podziałach. Jest ona równa $\binom{3}{2} + \binom{3}{2} = 6$. Druga n_2 liczba to liczba par, które należą do skupienia u_i ale nie należą do v_i. Analogicznie określamy liczbę n_3. Te dwie wartości są odpowiednio równe: $\binom{4}{2}+\binom{2}{2}+\binom{4}{2}-6 = 6+1+6-6 = 7$ oraz $\binom{3}{2} + \binom{3}{2} + \binom{4}{2} - 6 = 3 + 3 + 6 - 6 = 6$. Z kolei n_4 reprezentuje liczbę par, które w obu podziałach nie należą do tych samych skupień. Jest ona równa $\binom{10}{2} - 6 - 7 - 6 = 45 - 19 = 26$. Zatem otrzymujemy kolejne wartości indeksów:

$$RI = \frac{6 + 26}{45} = 0,711$$
$$JI = \frac{6}{6 + 7 + 6} = 0,315$$
$$RI_{popr} = \frac{15 - \frac{21 \cdot 15}{45}}{\frac{1}{2}(21 + 15) - \frac{21 \cdot 15}{45}} = \frac{8}{11} = 0,727$$

Klasyfikacja metod i algorytmów

W zasadzie trudno wskazać jeden właściwy podział metod i algorytmów analizy skupień. Dzieje się tak z uwagi na dużą liczbę kryteriów wedle których możemy dokonać ich klasyfikacji. W literaturze możemy spotkać się z różnymi klasyfikacjami metod analizy skupień (zob. [24] oraz [32]). Niektóre z nich, jak np. w [6] zdają się nie rozróżniać metod od algorytmów proponując kryteria o różnych stopniu szczegółowości (np. w jednym miejscu pyta, czy metoda jest hierarchiczna, a w innym, czy algorytm zakłada ustaloną liczbę skupień.).

Przedstawiona poniżej klasyfikacja jest nawiązaniem do klasyfikacji metod skalowania zilustrowanej w Rozdziale 1. Mowiąc o różnych metodach skupimy się na trzech najważniejszych kryteriach:

1. Jaki charakter przyporządkowania obserwacji do klas? (deterministyczny lub probabilistyczny)?

2. Jaki jest kształt macierzy podziału? (ostry vs. rozmyty)

3. Jaka jest relacja między skupieniami? (hierarchiczna vs. jedno poziomowa)

Ad. 1 Związek obserwacji ze skupieniami ma charakter probabilistyczny, jeśli na przestrzeni skupień można określić rozkład prawdopodobieństwa zmiennej określającej przynależność obserwacji do danego skupienia. Innym słowy, dla każdej obserwacji można wyznaczyć prawdopodobieństwa znalezienia się w danym segmencie. Jest ono wyznaczane w procesie estymacji parametrów modelu. Do takiego podejścia niezbędne są odpowiednie założenia na temat losowego charakteru próby i natury rzeczywistości.

Wiele klasycznych metod i algorytmów analizy skupień ma charakter deterministyczny. Ich przynależność do skupień jest efektem pewnych procedur optymalizacyjnych. Mimo optymalności uzyskiwanych podziałów, na ogół nie jest znane ryzyko związane z błędną klasyfikacją. Każda obserwacja należy do jednego i tylko jednego skupienia z prawdopodobieństwem równym 1.

Ad. 2 Jesteśmy przyzwyczajeni do myślenia o klasyfikacji, jako o pewnej funkcji ze zbioru obserwacji w zbiór skupień tj. że każdemu obiektowi przyporządkowane jest dokładnie jedno skupienie. Jest to zgodne z przyjętą definicją podziału, że składa się on ze zbiorów wzajemnie rozłącznych. Macierz podziału w klasycznej wersji składa się z elementów będących zerami lub jedynkami. Tak definiowany podział nazywać będziemy **ostrym** (ang. hard) i występuje on w przeważającej większości znanych nam metod.

Gdyby jednak zachować tylko pewne warunki brzegowe macierzy podziału (suma w każdej kolumnie jest równa 1, suma elementów w wierszu jest nie większa od liczebności całego zbioru) i osłabić „zero-jedynkowy" warunek na wartości jej elementów, wówczas uzyskalibyśmy podział **rozmyty** (ang. fuzzy). Taka macierz wypełniona jest liczbami z przedziału $(0, 1)$, które spełniają wcześniejsze warunki brzegowe. [3]

W przeciwieństwie do podziału ostrego, obserwacja nie jest traktowana jako niepodzielny obiekt, lecz przeciwnie - może się "rozszczepić" i egzystować w kilku skupieniach jednocześnie z różnych stopniem natężenia. Podział rozmyty zajmuje szczególne miejsce w analizie skupień i pozostałych zagadnieniach klasyfikacyjnych, gdyż dla każdego obiektu automatycznie definiowana jest jego **funkcja przynależności**. Sam zbiór rozmyty A w przestrzeni X definiuje się jako zbiór uporządkowanych par:

$$A = \{(x, \upsilon_X(x)) : x \in X\} \qquad , \upsilon_X : X \to (0, 1)$$

Praktycznie wszystkie kryteria optymalizacyjne zdefiniowane dla klasycznych podziałów przenoszą się na grunt klasyfikacji rozmytej z dokładnością do współczynnika przynależności. Dla przykładu odpowiednikiem algorytmu K-średnich jest w tym przypadku **Fuzzy C-means**, którego jedną z odmian jest implementacja w pakiecie R pod nazwą **fanny** (od ang. fuzzy analysis clustering). O ile K-średnich dążył do minimalizacji kryterium WSS, to jego rozmyty odpowiednik będzie szukał minimum poniższego wyrażenia:

$$\frac{\sum_{j=1,2...,k} \sum_{i,s} \upsilon(i,j)\upsilon(s,j)d(i,s)}{2\sum_i \upsilon(i,j)} \tag{2.9}$$

[3] Podział rozmyty jest uogólnieniem klasycznego podziału, podobnie jak cała logika rozmyta jest uogólnieniem klasycznej teorii zbiorów. Podstawy logiki rozmytej rozstały sformalizowane na początku lat '60 przez Zadeha, a przystępne wprowadzenie do teorii można znaleźć w [33].

Obecność logiki rozmytej w codziennym doświadczeniu badawczym zauważa P.Arabie [4, s.311] w pracy na temat zazębiających się skupień (ang. overlapping clusters) stwierdza, że klasyfikacja obiektów we wzajemne rozłączne i wyczerpujące cały zbiór segmenty, mimo, że jest metodologicznie elegancka, jest wątpliwa pod względem konceptualnym. Jako przykład, autorzy podają segmentację marek i konsumentów. Te pierwsze mogą jednocześnie być różnie pozycjonowane (np. guma do żucia może konkurować zarówno na rynku słodyczy jak i środków pielęgnujących jamę ustną), podobnie ich nabywcy mogą należeć do różnych segmentów jednocześnie (w końcu nikt nie lubi być szufladkowany i nie jest przywiązany tylko do jednego typu produktów).

Postać definicji zbioru rozmytego budzi skojarzenie z analizą klasy ukrytej Lazarsfelda, a dokładniej z funkcją śladu (traceline) lub z funkcją prawdopodobieństwa. W każdym przypadku dziedzina i zbiór wartości pokrywają się. Główna różnica polega na interpretacji, choć jej wyjaśnienie znacznie wykraczałoby poza ramy tej pracy. [4]. Bardziej powinniśmy skupić się raczej na konsekwencjach podejścia niż na teoretycznych rozważaniach. Dlatego rozsądnym wydaje zgodzić się ze stwierdzeniem Zadeha, że *logika rozmyta to zamaskowana teoria prawdopodobieństwa* (ang. probability theory in disguise) i że należy na nią patrzeć jako na teorię uzupełniają a nie konkurencyjną wobec prawdopodobieństwa.

Ad 3. Ostatnim kryterium jest relacja między skupieniami. Analogiczne pytania w znanych nam metodach skalowania brzmiałyby np. „Czy model zakłada liniowe skorelowanie czynników?" lub „Jaka jest postać relacji między blokami podziału?". Na relację między skupieniami możemy patrzeć dwojako. Po pierwsze, „zależność" skupień może być rozumiana jako wzajemne nachodzenie na siebie. Z tą sytuacją mieliśmy do czynienia, gdy miały one niepuste przecięcie tj. gdy uzyskany podział był rozmyty. Druga perspektywa porusza kwestię zawierania się skupień w zależności od przyjętej miary podobieństwa.

Możemy np. założyć, że obserwacje tworzą strukturę podobną do taksonomii żywych organizmów. Im większy poziom ogólności lub abstrakcji danego pojęcia, tym obserwacje są bardziej podobne. Oznacza to, że tworzą niewiele grup o stosunkowo dużych liczebnościach. W miarę zwiększania stopnia szczegółowości każda z grup dzieli się na mniejsze podgrupy, aż do uzyskania pojedynczych organizmów. Taki schemat grupowania nazywa się **hierarchicznym**.

Natomiast, gdy nie wprowadzamy żadnych dodatkowych założeń, mamy do czynienia z grupowaniem **nie hierarchicznym** zwanym również *obszarowym* lub *jednego poziomu* (ang. single level, [6]).

Tabela 2.4: Klasyfikacja metod analizy skupień

typ relacji	podział	hierarchiczna
deterministyczny	ostry	taksonomia wrocławska, metody SAHN
	rozmyty	xxx
probabilistyczny	ostry	grupowanie dwustopniowe (TwoStep)
	rozmyty	xxx

Podstawowe zasady działania najpopularniejszych algorytmów zostały szczegółowo opisane w Aneksie. Warto zapoznać się z nimi już w tym momencie, ponieważ w kolejnej sekcji będziemy się bezpośrednio odwoływać do ich właściwości.

[4] Przykłady niekończących się dyskusji (np. http://www.dbai.tuwien.ac.at/marchives/fuzzy-mail96/0166.html) mogą wyglądać zniechęcająco

2.2. Klasyczna analiza skupień w świetle postulatów dobrej metody

Mówiąc o klasycznej analizie skupień mamy na myśli zbiór tych metod, które nie wykraczają poza eksploracyjny i czysto opisowy charakter analizy skupień. W większości przypadków w ogóle nie dotykają one problemów omawianych w dalszej części. W momencie ich wynalezienia oraz implementacji nie myślano o nich jako o metodach skalowania. Rachunek prawdopodobieństwa czy elementy wnioskowania statystycznego były w nich nieobecne. Stopniowo zaczęto dostrzegać brak niektórych potrzebnych elementów. Zaowocowało to usprawnieniem wielu klasycznych metod i algorytmów, jednak sposób myślenia o analizie skupień nie zmienił się. Nadal postrzegana jest ona jako narzędzie *ad hoc*. W rezultacie, wiele praktycznych usprawnień, które pojawią się w kolejnych sekcjach może sprawiać wrażenie "łatania dziur", wynikających z braku wcześniejszego systematycznego myślenia o analizie skupień.

2.2.1. Segmentowalność zbioru

> Główny nacisk (w analizie skupień) nadal położony jest na szczegóły techniczne niż na testowanie dotyczące istnienia skupień [7]

Zaczynamy od pojęcia segmentowalności, gdyż z zgodnie z intuicją, pytanie o to, czy w zbiorze można wyróżnić jakąś strukturę powinno być zadane we wczesnym etapie analizy skupień. Podobnie postępujemy przeprowadzając analizę czynnikową - wielu pakietów statystycznych oferuje precyzyjne wskaźniki zdające sprawę ze struktury skorelowania między obserwowalnymi zmiennymi. Takimi syntetycznymi miarami są np. test Kaisera-Meyera-Olkina, wyznacznik macierzy korelacji czy macierz przeciwobrazów korelacji cząstkowych. Jednak w gruncie rzeczy podstawą wszystkich tych miar jest macierz korelacji lub kowariancji zmiennych.

Jeśli chodzi o analizę skupień, to jak zauważą Gordon [29, s.128]: Nie ma jednego oczywistego sposobu określania obecności (lub jej braku) jakiejś struktury. Problem polega jednak na tym, że o istnieniu segmentów analityk dowiaduje się dopiero po przeprowadzeniu analizy skupień. Ponadto informacja ta może być obarczona błędem związanym z wyborem danej metody analizy, który to błąd jest trudny do oszacowania.

Pozostaje pytanie, czy taka miara jest w ogóle niezbędna?

Everitt [24, s.180-181](?) twierdzi, że taki test nie musi być konieczny, pod warunkiem, że kierujemy się względami praktycznymi, tzn. z góry zakładamy pewną liczbą skupień i chcemy tylko przetestować kilka koncepcji. Podkreśla jednak dalej, że jeśli naszym celem jest wykrycie nieznanej struktury to zastosowanie testu jest kluczowe.

Podobnego zdania jest Bock [14, s.77]. Zastosowanie danego algorytmu analizy skupień zawsze doprowadzi nas do jakiegoś rozwiązania bez względu na to, czy w zbiorze jest struktura czy też nie. Nie jest to problem, gdy z góry znamy postać zakładanej finalnej struktury. Jednocześnie sygnalizuje, że oprócz tradycyjnych (opisowych i graficznych) metod eksploracyjnych, badanie struktury powinno odbywać się za pomocą modeli probabilistycznych i odpowiednich testów istotności statystycznej.

Do budowy takiego testu niezbędne są dwa elementy: pewien **model** opisujący powstawanie danych przy założeniu braku struktury oraz pewna **statystyka** testowa służąca do pomiaru odstępstwa (ang. departures) rozkładu empirycznego od modelu amorficznego.

Poniżej przedstawimy dwie koncepcje rozwiązania problemu segmentowalności zbioru. Ich opis pochodzi z [29] i [14], a są one oparte są na pewnych modelach statystycznych. Pierwszy z

nich bazuje na założeniu jednomodalności rozkładu, drugi - na macierzy losowej podobieństwa (ang. random dissimilarity matrix).

Analiza skupień, jak zaznaczyliśmy wcześniej, zaczyna się od analizowania macierzy odległości między obserwacjami. Analogicznie do modelu czynnikowego, można pokusić się o stworzenie takiej syntetycznej miary opartej tylko na informacji na temat odległości, która pozwalałaby ocenić zasadność poszukiwania struktury w zbiorze.

Modele statystyczne

Ling [45, s.160] twierdzi, że można zaproponować test służący do weryfikowania hipotez na temat braku struktury w zbiorze. Podstawą takiego testu miałby być model zgodny z hipotezą zerową na temat rozkładu macierzy odległości między obiektami.

Jednak, jak wskazuje Sneath [56], już po bliższym wglądzie w specyfikę problemu, okazuje się, że nie jest łatwo sformułować rozsądne postaci hipotez. Podaje trzy możliwe powody:

1. Za pomocą jakiego rozkładu opisać rozkład skupień?
2. Jak skorelowanie między wskaźnikami może wpływać na rozkład skupień?
3. Jeśli mówimy o zupełnej losowości, to według jakiego rozkładu?

Częściowo ten problem został rozwiązany przez Bocka [14], który zaproponował uniwersalne sformułowania hipotez zerowych i alternatywnych. Warto podkreślić, że poniższe testy są niezależne od całej procedury analizy skupień, w tym sensie, ze mogą one być przeprowadzone przed zastosowaniem konkretnego algorytmu. O segmentowalności zbioru nie dowiadujemy się zatem *ex post* jak w przypadku większości metod klasycznych. Wyjątkiem jest hipoteza na temat optymalności podziału, ale o niej powiemy więcej w następnej sekcji.

Jeśli chodzi o statystyki testowe, to ich postać nie odbiega od zasadniczego schematu. Zazwyczaj będą one wiązać informacje: na temat średniego podobieństwa (średniej odległości) między obiektami i najmniejszego lub największego podobieństwa (największej lub najmniejszej odległości) między obiektami. Ich wartości obliczane są bezpośrednio z macierzy odległości. Dodatkowo, dla testowania hipotezy na temat optymalności podziału wykorzystują się zmodyfikowaną postać statystyki F.

Hipotezy zerowe Bock podaje dwie różne postaci hipotez zerowych, głoszących, że zbiór pozbawiony jest jakiejkolwiek struktury. Oczywiście takich sformułowań może być znacznie więcej. Tu ograniczymy się do dwóch najważniejszych.

Równomierność rozkładu Rozkładem równomiernym (równoważnie: jednostajnym) często posługujemy się, gdy mamy na myśli "losowość". Wówczas nie tyle zależy nam na doprecyzowaniu sensu owej "losowości", co na podkreśleniu, że żaden wynik nie jest bardziej "prawdopodobny" od pozostałych. Gęstość jednowymiarowego (na prostej) rozkładu jednostajnego jest stała. Przypadek ten łatwo uogólnić na rozkład na kwadracie i na dowolnej p -wymiarowej kostce (kuli). Dla ogólnego przypadku możemy zapisać postać hipotezy zerowej:

$$f(x) = f_0(x) := \mathbf{1}_G(x)\frac{1}{vol(G)} \tag{2.10}$$

Gdzie $f(x)$ oznacza gęstość rozkładu profilu obserwacji, a $vol(G)$ jest p- wymiarową objętością pewnego obszaru. Wykresy rozrzutu generowane przez rozkłady jednostajne nie są skoncentrowane wokół żadnego punktu. W wyniku nieskończonego procesu losowania punktów o danej gęstości uzyskamy całkowite pokrycie obszaru, na którym rozpatrujemy rozkład.

Jednomodalność rozkładu Intuicja, która stoi za jednomodalnością rozkładu jest następująca. Otóż, jeśli w danej populacji istnieje wartość, która pojawia się najczęściej (wartość modalna), to wartości w wylosowanej próbie nie powinny "znacząco" odchylać się od tej wartości. Z formalnego punktu widzenia, wartość modalna jest wyznaczana w punkcie, w którym funkcja gęstości przyjmuje maksimum. Hipoteza o jednomodalności głosi, że takie maksimum jest jedynym maksimum globalnym i nie istnieją różne od niego maksima lokalne. Przykładami takich rozkładów są rozkład normalny, t-studenta lub chi-kwadrat. Ze względu na skomplikowany zapis hipotezy zerowej, ograniczymy się do wskazania, że $f_0(x) := h(x)$, gdzie $h(x)$ jest taką funkcją gęstości, że istnieje dokładnie jeden taki punkt y, że w jego otoczeniu funkcja ta zmienia znak. Wykresy rozrzutu generowane przez rozkłady jednomodalne nie są rozproszone, jak w przypadku równomiernego rozkładu, lecz są skoncentrowane wokół wybranej wartości. Przykładem realizacji takiego rozkładu jest amorficzny **zbiór** A.

Hipotezy alternatywne Hipotezą alternatywną dla równomierności rozkładu jest po prostu hipoteza głosząca, że zmienna ma dowolny inny rozkład, który przejawia pewne nieregularności. Można zatem zapytać, czy przedstawione powyżej hipotezy zerowe nie sobą względem siebie konkurencyjne. W końcu pierwsza nich zakłada równomierność czyli wielomodalność rozkładu, gdzie wszystkie modalne mają jednakowe prawdopodobieństwo. Rozwiązanie tej pozornej sprzeczności polega na tym, że brak struktury można definiować na co najmniej dwa sposoby. Używając fizycznej interpretacji, **amorficzność w sensie jednostajności** oznacza, że *masa układu* obiektów nie jest skoncentrowana w żadnym punkcie. Natomiast amorficzność w sensie jednomodalności oznacza, że praktycznie cała masa skupiona jest w jednym punkcie. W obydwu przypadkach, alternatywą będzie sytuacja pośrednia tj. gdy łączny rozkład wskaźników posiadać będzie skończoną, znacznie mniejszą od liczebności zbioru liczbę modalnych.

Formalnie można to zapisać jako $f_1(x) := h(x)$, gdzie $h(x)$ przyjmuje lokalne maksimum dla skończonej liczby punktów $y_1, y_2, ..., y_k$ i $k < n$.

Zauważmy, że jest to równoważne hipotezie, że łączny rozkład wskaźników jest generowany przez pewną kombinację rozkładów, co można zapisać jako:

$$f_1(x) := \sum_{j}^{j} \lambda_j \cdot f_j(x) \tag{2.11}$$

Takie sformułowanie problemu bliskie jest tzw. **skończonym mieszaninom rozkładów** (ang. finite mixture models), które omówimy dokładniej w następnym rozdziale.

Ling [45, s.160] uważa, że zaproponowane modele są realistyczne dla większości praktycznych problemów, jednak jedyną metodą poszukiwania odpowiednich statystyk testowych są metody symulacyjne (np. Monte Carlo). Mimo usilnych prób rozwiązania problemu metodami analitycznymi Ling doszedł do wniosku [45, s.162], że poza niektórymi prostymi przypadkami nie jest możliwe rozwinięcie teorii prawdopodobieństwa dla testowania hipotez zerowych, głoszących, że zbiór posiada określoną liczbę k skupień dla $k \geqslant 2$. W rozdziale 4, poświęconym w całości testowaniu hipotez na temat liczby skupień rozwiniemy tę problematykę i postaramy się opisać propozycję pewnych rozwiązań.

2.2.2. Miary dopasowania modelu do danych. Jakość podziału.

Choć na ogół pakiety statystyczne, w których zaimplementowane są klasyczne metody analizy skupień nie oferują żadnej miary uzyskanego podziału, to w literaturze można znaleźć

szeroką gamę takich miar. Niestety, jak zobaczymy poniżej, żadna z nich nie rozważa jakości podziału jako stopnia odtwarzalności rozkładu wskaźników. Z drugiej strony autorzy, którzy zajmowali się dokładnym testowaniem różnych kryteriów (zob. [50]) sami wskazują, że badacz często stoi w obliczu pytania, czy uzyskany podział odpowiada "rzeczywistym" partycjom, czy też został dokonany w sposób losowy. Jednocześnie podkreśla (s.187), że *potrzebna jest pewna **statystyka**, która zdawałaby sprawę ze stopnia rekonstrukcji empirycznej struktury zbioru.*

Krystalizacja struktury

Inspiracją do sposobu myślenia o mierze dopasowania modelu do danych jako o stopniu rekonstrukcji wskaźników może być współczynnik wykorzystywany w modelowaniu blokowym. Nosi on nazwę współczynnika **krystalizacji struktury** sieci. Pytanie o najlepszy możliwy podział zawiera się w pytaniu o istnienie podziału w ogólności, dlatego miara krystalizacji jest jednocześnie miarą segmentowalności zbioru.

Z naszego punktu widzenia najistotniejsze wydaje się pytanie: Czy w zbiorze istnieją bloki, a jeśli tak, to jaka permutacja obiektów najlepiej wyznacza podział na nie?

Jak podaje Banaszak [9]: wystarczającym argumentem przemawiającym za tezą o braku struktury przez sieć jest bezwartościowość wszelkiej **informacji** o podziale jej obiektów na pozycje. Bezwartościowość informacji oznacza jej zerową **wartość pragmatyczną**. W dalszej części czytamy, że: *Jest to równoważne stwierdzeniu, że brak struktury sieci oznacza, że jej rekonstrukcja za pomocą jednej liczby daje w efekcie taką samą wartość oczekiwaną straty jak rekonstrukcja wykorzystująca informację o podziale.*

Poprzez rekonstrukcję należy tutaj rozumieć poprawne przewidywanie elementu macierzy relacji między daną parą obiektów. Gdy jego predyktorem jest jedna liczba oznacza to, że niezależnie od wyznaczonego podziału, w każdym bloku struktura relacji wygląda identycznie jak w całej sieci. W związku z tym podział na bloki w ogóle nie poprawia przewidywania. Błąd przewidywania mierzony jest przy pomocy odpowiedniej **funkcji straty**. W tym przypadku przyjmuje się, analogicznie jak przy regresji średnich, kwadratową funkcję straty:

$$l(a, x_{ij}) = (x_{ij} - a)^2 \tag{2.12}$$

W tym wypadku a jest naszym przewidywaniem X - wartości relacji między dowolną parą obiektów. Wartością, która minimalizuje średnią wartość powyższego wyrażenia jest $g(X)$ - średnia liczba jedynek w sieci, która nazywana jest **gęstością sieci**. Formalnie:

$$g(X) = \frac{1}{n} \sum_{i}^{n} \sum_{j}^{n} x_{ij} \tag{2.13}$$

Zatem wartość funkcji straty, gdy nie wykorzystujemy informacji o podziale na bloki będzie równa wariancji gęstości sieci, czyli $g(X)(1 - g(X))$.

Zobaczmy, w jaki sposób wykorzystanie podziału może polepszyć nasze przewidywanie tj. zmniejszyć wartość funkcji straty. Zgodnie z intuicją, podział sieci na n bloków (gdy każdy obiekt jest dla siebie oddzielną klasą) powinien zapewnić minimalizację wariancji, jednak jest on mało praktyczny ze względu na zbyt dużą liczbę parametrów.

Załóżmy więc, że mamy podejrzenie, że w sieci można wyróżnić k bloków tj. istnieje k-członowa struktura sieci. Dla ilustracji, przyjmijmy, że $k = 2$ oraz znamy liczebności każdego z bloków. Będziemy poszukiwać optymalnego podziału wierzchołków B_2. Nietrudno zauważyć,

że funkcja straty przyjmuje minimum dla $g(X) = 1$ lub $g(X) = 0$ tzn. gdy bloki "wypełnione" są wyłącznie zerami lub jedynkami. Będziemy szukać więc takiej permutacji wierzchołków, która zapewni nam tego typu strukturę sieci.

Dla ilustracji weźmy następującą sieć relacyjną (2.5).

Tabela 2.5: Przykładowa wyjściowa sieć relacyjna. $g(X) = 0,5$

	1	2	3	4	5	6	7	8	9	10
1	1	1	1				1	1		1
2		1	1	1					1	
3	1		1	1	1		1			1
4		1		1					1	
5	1		1		1			1		1
6		1		1	1	1			1	
7			1		1		1	1		1
8	1		1		1		1	1		1
9		1	1	1		1			1	
10		1			1		1	1		1

Załóżmy, że interesują nas podział na dwa człony: B_1 liczący cztery oraz B_2 liczący sześć obiektów.

Biorąc pod uwagę zakładaną strukturę zbioru, wszystkich możliwych permutacji obiektów jest w tym przypadku $\binom{10}{4}\binom{6}{6} = \binom{10}{4} = 210$, czyli dokładnie tyle ile odpowiednich podzbiorów. W ogólnym przypadku k członów n elementowej sieci liczba permutacji jest równa $\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots 1$, gdzie n_i oznacza liczebność bloku o numerze i. Wybierzmy jedną z możliwych permutacji (2.6).

Tabela 2.6: Przykładowe losowe przyporządkowanie obiektów do bloków. k=2

	1	2	3	4	5	6	7	8	9	10
1	1	1	1				1	1		1
2		1	1	1					1	
3	1		1	1	1		1			1
4		1		1					1	
5	1		1		1			1		1
6		1		1	1	1			1	
7			1		1		1	1		1
8	1		1		1		1	1		1
9		1	1	1		1			1	
10		1			1		1	1		1

W **??** podsumowano lokalne gęstości w blokach w zależności od podziału. Ta macierz nosi nazwę **wzoru strukturalnego**. Widać, że średnia wariancja gęstości jest wyraźnie mniejsza w przypadku optymalnego podziału. Aby porównać pragmatyczną wartość obu podziałów wystarczy porównać proporcje wyjaśnionej wariancji gęstości sieci:

$$K = \frac{D^2(g(X)) - E(D^2(g(X|\mathcal{B}_2)))}{D^2(g(X))} \tag{2.14}$$

Redukcja błędu przewidywania jest równa odpowiednio:

Tabela 2.7: Optymalne przyporządkowanie obiektów do bloków. k=2

	2	4	6	9	1	3	5	7	8	10
2	1	1		1		1				
4	1	1		1						
6	1	1	1	1			1			
9	1	1	1	1		1				
1	1				1	1		1	1	1
3		1			1	1	1	1		1
5					1	1	1		1	1
7						1	1	1	1	1
8					1	1			1	1
10	1						1	1	1	1

Tabela 2.8: Wzór strukturalny dla podziału dwuczłonowego

losowy		optymalny	
0,688	0,333	0,875	0,125
0,458	0,556	0,125	0,778

$$K_1 = \frac{0,25 - 0,236}{0,25} = 0,056 \tag{2.15}$$

$$K_2 = \frac{0,25 - 0,132}{0,25} = 0,472 \tag{2.16}$$

Zauważmy, że interpretacja współczynnika K jest identyczna jak kwadratu stosunku korelacyjnego η^2. Im większy stosunek wariancji międzygrupowej do wewnątrzgrupowej, tym wyraźniejszy stopień krystalizacji struktury. Rolę zmiennej objaśnianej stanowi tutaj kształt sieci relacyjnej, natomiast zmienna objaśniająca składa się z pary $(X^*, \mathcal{B}^*)$, gdzie pierwszy element oznacza wzór strukturalny (gęstość podmacierzy), a drugi jest macierzą optymalnego podziału.

Im większy stopień krystalizacji struktury, tym większa dokładność w rekonstrukcji wyjściowej macierzy relacji. Rekonstrukcja odbywa się za pomocą zwykłego mnożenia macierzy podziału i wzoru strukturalnego. Poniższe równanie nosi nazwę **równania reprodukcyjnego sieci**:

$$\hat{X} = \mathcal{B}^* \cdot X^* \cdot \mathcal{B}^{*T} \tag{2.17}$$

Przy kwadratowej funkcji straty średni błąd przewidywania będzie równy $1 - K$, gdzie K jest współczynnikiem krystalizacji struktury sieci.

Kryteria dobrego podziału

Problematyka struktury sieci w modelowaniu blokowym miała na celu zaznaczenie decyzyjnego charakteru analizy skupień. Wybór odpowiedniego podziału łatwo tłumaczy się na język zysków (z informacji) i strat (z powodu niedokładnego przewidywania). Wymaga

zdefiniowania odpowiednich funkcji celu (błędu, straty), których wartości będą stanowić kryteria optymalności podziału. Milligan [49] wyróżnia dwa rodzaje kryteriów: **zewnętrzne** i **wewnętrzne**. Jak wynika z nazwy, pierwsza grupa zakłada istnienie zewnętrznego standardu, z którym porównywany jest wynik analizy skupień. Przykładem takiego kryterium jest wspomniany indeks Randa pod różnymi postaciami. Natomiast kryteria wewnętrzne oparte są wyłącznie na informacji pochodzącej z procesu analizy skupień. Zwykle stanowią integralną część procedur grupowania, co sprawia, że stosuje się je znacznie częściej. Można podzielić na je dwie grupy: **wariancyjne** oraz **korelacyjne**.

Kryteria wariancyjne Podobnie jak w sytuacji modelowania blokowego, problem optymalizacyjny opiera się o pewną funkcję celu. Jej argumentami są pewne przekształcenia macierzy **W** - kwadratów odległości obserwacji od średnich wewnątrzgrupowych i **B** - kwadratów odległości średnich wewnątrzgrupowych od średniej w całej populacji.

Najczęściej oblicza się **wyznaczniki** lub **ślady** tych macierzy. W tym drugim przypadku otrzymujemy sumę elementów na przekątnej co w literaturze spotyka się pod nazwą sumy kwadratów odległości od średniej. W przypadku macierzy **W** mamy oznaczenie WSS (ang. Within-Groups Sum of Squares), a dla macierzy **B** mamy BSS (ang. Between Groups Sum of Squares).

$$TSS = \sum_{i}^{n} (x_i - \mu)^T (x_i - \mu)$$

$$WSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \mu_j)^T (X_{ij} - \mu_j)$$

$$BSS = \sum_{j=1}^{k} (\mu_j - \mu)^T (\mu_j - \mu)$$

Teoretycznym uzasadnieniem stosowania powyższych kryteriów jest twierdzenie o dekompozycji całkowitej sumy kwadratów odchyleń od średniej: [5].

$$\mathbf{TSS} = \mathbf{BSS} + \mathbf{WSS} \tag{2.18}$$

Istnieje kilka możliwości skorzystania z powyższego wzoru do budowy kryterium. Najczęściej wykorzystuje się różnicę między $B - W$ lub stosunek $\frac{B}{W}$. W obu przypadkach większa wartość kryterium oznacza wyraźniejszy podział zbioru na segmenty.

Warto zauważyć, że przy odpowiednich założeniach (normalność, homoscedastyczność) iloraz $\frac{B}{W}$ jest statystyką używaną do testowania równości średnich w analizie wariancji i zgodnie z hipotezą zerową głoszącą, że cała próba pochodzi z jednej populacji ma **rozkład F Fishera-Snedecora** o $(n - G, G - 1)$ stopniach swobody.

Warto podkreślić, że statystyka F ma sens jedynie gdy przyporządkowanie obserwacji do klas jest losowe. Natomiast w naszym przypadku jest wręcz niepożądane, aby algorytm działał losowo. Chcemy aby maksymalizował wartość ilorazu, co powoduje, że wartość F jest w większości przypadków "istotnie statystycznie" różna od zera, ale nie bardzo wiadomo, co miałoby to oznaczać.

Podstawowym terminem, który obecny jest w większości kryteriów jest tzw. **separowalność** lub po prostu odległość między skupieniami. Dany podział uznawany jest za dobry

[5] Czasem w przypadku jednowymiarowym spotyka się pod nazwą Wielkiego Twierdzenia o Wariancji

w momencie, gdy odległość między skupieniami jest znacznie większa od odległość między obserwacjami w skupieniach.

(tutaj przykład jednowymiarowej miary).

Przypadek wielowymiarowy jest mniej trywialny, ale idea pozostaje tama (Friedman Rubin, 1161). Wówczas do testowania używa się tzw. λ Wilksa zdefiniowanej jako $\frac{detT}{detW}$. Wysokie wartości tego współczynnika świadczą o wyraźnych różnicach między skupieniami. Przekształcając to wyrażenie otrzymujemy kolejne kryterium:

$$\frac{\det T}{\det W} = \frac{\det W + \det B}{\det W} = \det(I + W^{-1}B) \tag{2.19}$$

Drugi składnik macierzowej sumy pod wyznacznikiem jest elementem tzw. **śladu Hotellinga** (ang. Hotelling's trace) znanym pod postacią $tr(W^{-1}B)$. Na podstawie twierdzeń z algebry liniowej ostatnie dwa współczynniki wykazują ciekawe właxności. Są one niezmiennicze względem przekształceń linowych.

Mimo atrakcyjności powyższych współczynników należy pamiętać o trzech podstawowych problemach z nimi związanych. Po pierwsze, punktem wyjścia do ich zastosowania jest macierz danych surowych. Z drugiej strony warunkiem koniecznym, aby móc policzyć wariancję (lub sumę kwadratów odchyleń) jest co najmniej przedziałowy charakter skali.

Znacznie bardziej niepokojącym jest brak znajomości rozkładu tych statystyk. Jest tak, ponieważ uzyskany podział nie jest efektem pewnego procesu losowego lecz efektem działania algorytmu, który dąży do **dobrego podziału**. W konsekwencji nie powinien nas interesować rozkład statystyk $\frac{\det T}{\det W}$ czy $tr(W^{-1}B)$ lecz ich *maksimum*. Trudności związane z analitycznym opisem takich rozkładów zmuszają nas do sięgnięcia po metody symulacyjne.

Ponadto powyższe kryteria mogą nieść sensowną informację, o ile oceniamy różne podziały, ale dla ustalonej liczby skupień. Ponieważ wszystkie opierają się o minimalizację wariancji wewnątrzgrupowej, podział na $k+1$ skupień nie może być gorszy od podziału na k skupień. Wpadając w tę **pułapkę monotoniczności** w końcu uzyskamy "najlepszy" możliwy podział na jednoelementowe skupienia.

Czy w jakiś sposób możemy liczyć na kompromis między optymalizacją kryterium i liczbą skupień? Więcej o tym zagadnieniu powiemy w następnej sekcji. Na razie jednak dokończmy klasyfikację kryteriów.

Kryteria korelacyjne Stosowane są różne typy korelacji, w zależności od wykorzystanej metody analizy skupień. Najprostszą miarą jest **współczynnik korelacji Pearsona** między macierzą odległości między obserwacjami, a macierzą przynależności do tego samego skupienia. Kryterium to nosi specjalną nazwę - **point biserial correlation**.

Intuicja, która stoi za tą miarą jest bardzo prosta. Po skończonym procesie analizy skupień każdej parze obiektów przypisujemy 0 lub 1 w zależności od faktu, czy należą lub nie do tego samego skupienia. Wynik ten przedstawiamy w macierzy przynależności, która ma ten kształt i wymiar co macierz odległości. Następnie obie macierze przekształcane są w wektory zmienny, między którymi wyznaczany jest współczynnik korelacji. Przyjmuje on wartości z przedziału $(-1, 1)$. Jego dodatnia wartość oznacza wystąpienie nieprawidłowości w klasyfikacji, ponieważ większa odległość między obserwacjami nie powinna sprzyjać ich klasyfikacji w ramach jednego skupienia. Dla przykładu, rozpatrzmy następujący prosty zbiór

Macierz odległości przyjmuje następującą postać:

Binarna macierz przynależności do skupień wygląda następująco. Można zauważyć, że odpowiada ona macierzy danych relacyjnych po optymalnej permutacji, gdyż można w niej wyróżnić bloki wypełnione tylko jedynkami lub tylko zerami.

Tabela 2.9: Przykładowy zbiór z wyraźnymi dwoma skupieniami

nr	1	2	3	4	5	6	7	8	9	10
x	1,06	1,26	1,33	1,38	1,43	5,61	5,85	6,54	5,17	5,64

Tabela 2.10: Wyjściowa macierz odległości euklidesowych

	1	2	3	4	5	6	7	8	9	10
1	0,00	0,20	0,27	0,32	0,37	4,55	4,79	5,48	4,10	4,58
2	0,20	0,00	0,06	0,12	0,17	4,35	4,59	5,28	3,90	4,38
3	0,27	0,06	0,00	0,06	0,10	4,28	4,52	5,21	3,84	4,31
4	0,32	0,12	0,06	0,00	0,05	4,23	4,46	5,16	3,78	4,25
5	0,37	0,17	0,10	0,05	0,00	4,18	4,42	5,11	3,74	4,21
6	4,55	4,35	4,28	4,23	4,18	0,00	0,24	0,93	0,44	0,03
7	4,79	4,59	4,52	4,46	4,42	0,24	0,00	0,69	0,68	0,21
8	5,48	5,28	5,21	5,16	5,11	0,93	0,69	0,00	1,38	0,90
9	4,10	3,90	3,84	3,78	3,74	0,44	0,68	1,38	0,00	0,47
10	4,58	4,38	4,31	4,25	4,21	0,03	0,21	0,90	0,47	0,00

Tabela 2.11: Macierz przynależności do skupień. 1 oznacza, że dwie obserwacje należą do tego samego skupienia, 0 - do różnych

	1	2	3	4	5	6	7	8	9	10
1	.									
2	1	.								
3	1	1	.							
4	1	1	1	.						
5	1	1	1	1	.					
6	0	0	0	0	0	.				
7	0	0	0	0	0	1	.			
8	0	0	0	0	0	1	1	.		
9	0	0	0	0	0	1	1	1	.	
10	0	0	0	0	0	1	1	1	1	.

Zauważmy, że powyższa macierz jest liniową transformacją uproszczonej postaci wyjściowej macierzy odległości. Oznacza to, że jeśli wartość współczynnika korelacji jest wysoka, to z dokładnością do przeskalowania i przesunięcia bloków obie macierze są identyczne. Kwadrat współczynnika korelacji liczony dla wektorów utworzonych z odpowiadających sobie elementów powyższych macierzy wynosi 0.958.

Powyższe kryterium jest szczególnym przypadkiem współczynnika korelacji rangowej. W tym miejscu rolę rang odgrywają odległości oraz przynależność do skupień. Dla każdej pary sprawdzana jest zgodność między poziomem odległości i zmienną zero-jedynkową. W rezultacie uzyskuje się w pewnym sensie naturalny współczynnik zgodności klasyfikacji, lecz ze względu na nominalny charakter drugiej zmienne traci się pewne informacje.

Bardziej dokładnym kryterium jest **indeks sylwetki** (ang. silhouette index) skupienia, którego autorami są Kaufmann i Rouseeuuw (przewodnik po R). Ma on następującą postać:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

W tym miejscu a_i oznacza średnią odległość obserwacji x_i do wszystkich obserwacji w *tym samym* skupieniu. b_i z kolei oznacza średnią odległość obserwacji x_i do najbliższego ze skupień. Najbliższe skupienie definiowane jest jako takie, którego suma odległości od x_i jest najmniejsza. Indeks sylwetki obliczany jest dla każdej obserwacji, co pozwala na uśrednianie go na poziomie skupień czy całego zbioru. Wysokie (bliskie jedności) wartości współczynnika świadczą o poprawnym zaklasyfikowaniu. Obserwacje, dla których indeks waha się w okolicach zera znajdują się na granicy dwóch skupień, a ujemne wartości przemawiają za niepoprawną klasyfikacją (por. manual do R).

Specyficzną miarą dla metod hierarchicznych jest tzw. **współczynnik korelacji kofenetycznej** (ang. cophetenic correlation coefficient). Podobnie jak indeks sylwetki opisuje globalną relację między odległością obserwacji i ich klasyfikacją, przy czym ta ostatnia nie ma binarnego charakteru lecz wyraża się tzw. **współczynnikiem połączenia** (ang. fusion coefficient). Współczynnik ten jest podstawą konstrukcji **dendrogramu** (zob. aneks) i oznacza progową odległość, dla której dane dwa obiekty włączone zostały w ramach jednego skupienia. Im większa odległość między obserwacjami, tym większy spodziewany współczynnik fuzji (wysokość drzewa). Współczynnik korelacji kofenetycznej zdaje zatem sprawę z tego, w jakim stopniu relacje odległości między obserwacjami zachowywane są przez relacje wynikające ze struktury dendrogramu.

Korelacja kofenetyczna tworzona jest w analogiczny sposób, co opisana wcześniej point-biserial correlation. Elementy macierzy odległości i macierzy niosącej informację na temat wysokości drzewa przekształcane są w wektory, między którymi obliczany jest stopień zgodności.

Przykładowe wartości współczynników dla

2.2.3. Liczba skupień

> Wybór kryterium i algorytmu optymalizacyjnego nie wyczerpują problematyki analizy skupień. Cały czas musimy mieć na uwadze ogromny problem wyboru odpowiedniej liczby grup. [22]

Znaczenie problemu Przypomnijmy, że podstawowym zadaniem analizy skupień jest podział zbioru obiektów na grupy maksymalnie homogeniczne wewnątrz i maksymalnie heterogeniczne między sobą. Jednak tak sformułowany problem może nie wystarczyć do znale-

zienia optymalnego wyniku. Albo inaczej, w świetle dostępnych i zaprezentowanych powyżej kryteriów optymalności podziału jedynym optymalnym rozwiązaniem jest podział zbioru na pojedyncze obserwacje. Każda pojedyncza obserwacja reprezentuje pozbiór będący singletonem, a zbiór singletonów wyznacza prawidłowy podział zbioru. Dodatkowo, suma kwadratów odchyleń wewnątrzgrupowych jest równa zero, co utwierdza nas w przekonaniu, że uzyskany podział jest najlepszy.

Jednak tak rozwiązany problem nie jest zgodny z ideą skalowania czy też modelowania w ogólności. Opisanie zbiorowości za pomocą parametrów, których liczb jest równa liczbie obserwacji jest ze statystycznego punktu widzenia bezużyteczne. Sztuka polega na tym, aby opisać populację za pomocą możliwie najmniejszej liczby parametrów i uzyskać wysoki stopień dokładności. Te dwie cechy dobrego modelowania nazywane są odpowiednio prostotą (ang. parsimony) i dopasowaniem do danych (ang. goodness of fit).

W analizie skupień, prostota modelu ma dwojaki sens. Po pierwsze, chodzi o prostotę parametrów opisujących skupienia. Bez względu, czy to będzie wektor średnich (centroid), czy średnica (największa odległość między obserwacjami w skupieniach), to każda z tych parametryzacji jest lepsza (bo prostsza) od wypisania wszystkich obserwacji i ich profili w każdym skupieniu. Drugim aspektem prostoty jest liczba skalowanych wartości klasy ukrytej, czyli liczba skupień. Poświęcimy mu najwięcej miejsca, nie tylko dlatego, że całkowita liczba parametrów bezpośrednio zależy od liczy klas. Zagadnienie to jest ściśle powiązane z pozostałymi podstawowymi problemami analizy skupień.

Pozytywna odpowiedź na pytanie, czy zbiór jest segmentowalny rokuje na wyznaczenie dokładnej liczby skupień. Z drugiej strony, zaczynając od identyfikacji ich liczby natychmiastowo uzyskujemy odpowiedź na pytanie, czy zbiór posiada jakąkolwiek strukturę.

Podobnie, ze zbioru optymalnych rozwiązań wybierzemy wynik z najmniejszą liczbą skupień. Z drugiej strony, bez wcześniejszego ustalenia ich liczby problem optymalizacyjny może być źle postawiony.

Wyznaczenie odpowiedniej liczby skupień zbliża nas do podjęcia odpowiedniej decyzji, co do jednostek odstających. Z drugiej strony ich identyfikacja wyklucza przypadki, gdy uwzględniane będą mało liczne, zakłócające strukturę zbioru, skupienia.

W niektórych przypadkach, informacja na temat liczby skupień podyktowana jest względami teoretycznymi lub praktycznymi. W pierwszym przypadku możemy weryfikować, czy w zbiorze faktycznie istnieje zakładana liczba skupień lub optymalizować podział w ramach założeń tej teorii. W drugim przypadku, liczba skupień determinowana jest przez zakładane cele badawcze lub strategiczne np. gdy zależy nam na wyróżnieniu określonej a priori liczby segmentów.

Dlatego warto podkreślić, że w dalszej części interesować nas będzie analiza skupień bardziej jako metoda eksploracji danych i poszukiwania struktury. Taka perspektywa bliższa jest oryginalnej definicji analizy skupień, mówiącej, że dysponujemy minimum informacji tj. wyłącznie rozkładem łącznym wskaźników.

Problem liczby skupień nie pozostaje w izolacji względem pozostałych zagadnień. Mimo to, nie ma zgody, co do uniwersalnego kryterium wyznaczania ich odpowiedniej liczby. Niektórzy autorzy, jak podaje Bailey poddają w wątpliwość możliwość ich wyznaczenia i uznają, że priorytetem powinna być prostota oraz interpretowalność rozwiązania, gdyż każda ścisła miara jakości dopasowania jest bezproduktywna ([6]).

Nie należy tego rozumieć jako oznaki zmęczenia i rezygnacji z poszukiwań takiej miary. Wręcz przeciwnie, od początku lat 70-tych powstało wiele różnych kryteriów. Omówimy tylko wybrane z nich.

Na początku należy przypomnieć słowa Fraley i Raftery [25], ze żadna z metod hierarchicznych czy relokacyjnych nie porusza bezpośrednio problemu liczby klas w rozumieniu

statystycznym. Do podobnego wniosku dochodzą Milligan i Cooper [50]: *W rzeczywistości, żadna z procedur analizy skupień nie dostarcza w ogóle lub jedynie w małym stopniu informację na temat liczby skupień w zbiorze.*

Tego samego zdania są twórcy SPSS: *Żadna z metod analizy skupień nie dotyka bezpośrednio problemu określania liczby skupień, ponieważ środki służące do tego celu są skomplikowane i zwykle jest on rozpatrywany oddzielnie. Jest wiele strategii, którymi można się posłużyć do identyfikacji liczby skupień.* [57]

Przegląd kryteriów Czytelnik, który zainteresowany jest wyczerpującym i szczegółowym przeglądem najbardziej popularnych kryteriów powinien zapoznać się z tekstem ([50]). Artykuł ten jest próbą systematycznego porównania 30 miar, wśród których można znaleźć zarówno skrajnie heurystyczne, jak i te wyrafinowane pod względem matematycznym. Większość opisanych kryteriów można pogrupować w szersze kategorie (analogicznie jak uczyniliśmy to z kryteriami optymalnego podziału). Dodatkowo, zaprezentujemy najnowsze pomysły pochodzące z innych źródeł.

W ogólnym przypadku, kryterium optymalnej liczby skupień powinno zawierać dwa elementy: miarę optymalnego podziału oraz odpowiadającą jej liczbę skupień. Kryteria te zatem można łatwo skonstruować poprzez dołączenie informacji na temat liczby klas dla wyliczonej wartości kryterium. Ogólnie rzecz biorąc zadanie polega na znalezieniu pewnego rodzaju kompromisu między prostotą modelu, a jakością dopasowania modelu do danych wyrażoną poprzez pewną funkcję celu.

Najprostszym przykładem kryterium jest wzbogacenie składnika WSS o czynnik k, gdzie k jest liczbą skupień. Marriott ([50]) proponuje wybór $k^2 \cdot WSS$. Ciekawe uzasadnienie, dlaczego występuje k w drugiej potędze można znaleźć w [**?**]. Zachowanie się współczynnika jasno wynika z tego postaci: wzrost liczby skupień powoduje spadek wartości śladu macierzy. Niekiedy ten czynnik nosi nazwę funkcji kary (ang. penalizing factor). Zakładamy, że w przypadku obecności wyraźnej struktury zbioru, od pewnego miejsca wzrost liczby klas będzie powodował coraz wolniejszy spadek sumy kwadratów wewnątrz grup, co oznaczać będzie wzrost całego kryterium. Optymalna liczba skupień zdefiniowana jest zatem jako k, które minimalizuje wartość $k^2 \cdot WSS$.

Wizualnym analogiem tego kryterium jest kopia znanego z analizy czynników lub analizy głównych składowych **wykresu osypiska** (ang. scree plot lub Cattell's plot), który przedstawia zależność sumy kwadratów wewnątrzgrupowych względem liczby skupień. Przy założeniu, że dla ustalonej liczby skupień algorytm znajduje optymalne rozwiązanie, punkty na wykresie powinny być w przybliżeniu ułożone wzdłuż hiperboli:

(tu rysunek)

Widzimy zatem, że w miarę kolejnych podziałów zbioru na coraz większą liczbę klas kryterium maleje, jednak dzieje się to co raz "wolniej". Innymi słowy, maleje użyteczność kolejnych skupień, a rośnie koszt związany z rozbudową modelu o dodatkowe parametry. Ta prosta heurystyka została przekształcona przez Tibshiraniego i in. ([**?**]) w model statystyczny. Idea polega na standaryzacji wykresu osypiska i odniesieniu go do pewnego wykresu zgodnego z hipotezą zerową. Ogólna postać statystyki dla próby o liczebności n i k skupieniach ma następującą postać:

$$Gap_{n,k} = E(\log W_k) - \log W_k \tag{2.20}$$

Statystyka ta nosi nazwę **statystyki odstępu** (ang. gap statistic) ponieważ mierzy punktowe różnice między modelem amofircznym, a empirycznym. Jak pamiętamy, jedną z możliwych hipotez zerowych na temat amorficzności zbioru jest hipoteza o jednostajności rozkładu

po współrzędnych. Autorzy wyliczyli przybliżoną wartość oczekiwaną $E(\log W_k)$ dla p zmiennych z dokładnością do stałej:

$$E_{k,p,n} = \log\left(\frac{pn}{12}\right) - \frac{2}{p}\log k \tag{2.21}$$

O ile w przypadku wykresu osypiska interesowała nas liczba skupień, dla których następował wyraźny spadek wartości kryterium, o tyle tutaj wskazówką jest największa różnica między wartością oczekiwaną statystyki, a empiryczną wartością logarytmu sumy kwadratów odchyleń wewnątrzgrupowych.

(tu też rysunek z tekstu)

Inną modyfikacją kryterium $k^2 \cdot WSS$ jest **indeks Krzanowskiego-Lai**, który ukazuje różnice między kolejnymi sumami kwadratów biorąc pod uwagę również wymiar przestrzeni. Z postaci indeksu wynika, że wraz ze wzrostem p wymiarów przestrzeni rośnie znaczenie współczynników W_k.

$$KL(k) = \frac{|DIFF(k)|}{|DIFF(k+1)|} \tag{2.22}$$

$$DIFF(k) = (k-1)^{\frac{2}{p}} W_{k-1} - k^{\frac{2}{p}} W_k \tag{2.23}$$

Podobnie jak we wcześniejszym przypadku, należy wybrać takie k, które maksymalizuje wartość wyrażenia KL tj. wyznacza największą wartość „skoku" dla kolejnych skupień.

Należy zachować ostrożność przy szacowaniu kolejnych wartości współczynników, niezależnie od tego, czy bazujemy wyłącznie na wykresie czy konkretnych obliczeniach. Dowód na to, jak bardzo nieczuła może być nasza intuicja został przedstawiony w pracy Sugar i James [**?**]. Podają oni przykład trzech różnych zbiorów, w których występują odpowiednio jedno, trzy i sześć skupień. Na każdym z nich przeprowadzono metodę K-średnich dla różnej liczby skupień od 1 do 10 i sporządzono wykres zależności miary rozproszenia [6] (ang. distortion) w zależności od liczby skupień. Każdy z nich miał ten sam hiperboliczny kształt. Przykład miał na celu pokazanie, jak dalece możemy mylić się wykorzystując kryteria wizualne, które ignorują wymiarowość przestrzeni.

Przyczyną nieprawidłowej identyfikacji było założenie o stałej postaci wykresu niezależnie od wymiaru przestrzeni. Naturalnym rozwiązaniem problemu była zatem odpowiednia transformacja miary rozproszenia. Bazując na twierdzeniach z tzw. teorii sygnałów będącej częścią ogólnej teorii informacji, autorki dochodzą do wniosku, że modelowa zależność między liczbą skupień a miarą rozrzutu uzależniona jest od wymiaru przestrzeni i wyraża się następująco:

$$d_k = k^{-\frac{2}{p}}$$

Kryteria wariancyjne optymalności podziału doprowadziły do powstania pokrewnych wskaźników dla liczby skupień. W przypadku niektórych, modyfikacja ogranicza się jedynie do zmiany jednego znaku drukarskiego (poprzez uzależnienie kryterium od k tj. dopisaniu odpowiedniego indeksu). Tak jest np. w przypadku indeksu Friedmana i Rubina, który należy zmaksymalizować względem k:

[6] Standardowo za miarę rozproszenia przyjmuje się sumę kwadratów odchyleń wewnątrz grupowych - W. W artykule zamiast niej występuje średnia odległość Mahalanobisa uwzględniająca skorelowanie między wskaźnikami w skupieniach

$$FR(k) = \frac{\det T}{\det W_k} = \det W_k^{-1} B$$

Z kolei Celiński i Harabasz stworzyli indeks, który wykazuje mocne podobieństwo do statystyki F

$$CH(k) = \frac{B_k}{W_k} \frac{n-k}{k-1}$$

Wartość k dla którego iloraz jest największy sugeruje odpowiednią liczbę skupień. Należy jednak pamiętać, że powyższe kryterium nie jest tożsame z testem F stosowanym między innymi w analizie wariancji. Tam, zgodnie z hipotezą zerową, iloraz sum kwadratów (z dokładnością do stałej) ma rozkład F-Snedecora, podczas gdy w analizie skupień wartość tego ilorazu jest maksymalizowana, wobec czego należy pytać nie o rozkład samego F, ale $\max F$.

Ciekawą modyfikacją pseudotestu F jest **indeks Hartigana**, który przypomina cząstkowy test F [**?**, s.758].

$$k = min\{j : (n-k-1)\left(\frac{W_j}{W_{j+1}} - 1\right) \leqslant 10\}$$

Podobnie, jak w przypadku wcześniejszych kryteriów sprawdza się użyteczność kolejnego $k+1$ skupienia w porównaniu z istniejącym podziałem na k. Wybierana jest najmniejsza wartość k, która zapewnia, że wartość indeksu będzie nie większa niż 10. Jak podaje Everitt [24] stała ta jest efektem badań symulacyjnych.

Indeksy związane z kryteriami wariancyjnymi na ogół związane są z metodami opartymi na relokacji obiektów (np k-średnich). Wiele z wykorzystywanych algorytmów ma zakodowane w swojej definicji optymalizację kryterium jednorodności skupień, co pozwala niemal natychmiastowo zaimplementować odpowiednie kryteria. Z drugiej strony, praktyka pokazuje, że ich obecność w standardowym oprogramowaniu (np. SPSS, Stata, R) jest rzadkim zjawiskiem.

Niestety, podobnie jest w przypadku kryteriów dla metod hierarchicznych. Przypomnijmy, że specyfiką tych metod jest tworzenie skupień w ramach pewnej struktury, a efektem końcowym jest jej wizualizacja w postaci drzewa czyli dendrogramu. Problem optymalnej liczby skupień sprowadza się do znalezienia punktu optymalnego ucięcia drzewa.

Podobnie jak w przypadku wykresu osypiska, identyfikacja tego punktu może odbyć się drogą zwyczajnej obserwacji. Pewną formalizacją tej metody jest tzw. **kryterium stepsize** wymyślone przez S.C. Johnsona. Polega ono na badaniu dynamiki współczynnika fuzji. Jeśli stosunki kolejnych współczynników zaczynają wyraźnie rosnąć, oznacza to, że istnieje wyraźna luka między obserwacjami lub skupieniami, co sugeruje wysoki stopień izolacji.

(przykład z R)

Próbą ujęcia problemu w kategoriach statystycznych jest **indeks Mojeny**, który również posługuje się współczynnikiem fuzji. Idea polega na potraktowaniu go jako zmiennej losowej o odpowiednim rozkładzie. Estymatory parametrów (wartości oczekiwanej i odchylenia standardowego) oblicza się bezpośrednio z wartości współczynników fuzji w otrzymanym dendrogramie. Mojena zakłada, że do pewnego miejsca proces skupiania odbywa się "regularnie" z dokładnością do losowego odchylenia, co objawia się równomiernym rozkładem kolejnych współczynników fuzji. W jednej z wersji, modelowym rozkładem był rozkład normalny, choć na ogół empiria nie potwierdza tego założenia. (przykład). Zadaniem indeksu Mojeny jest

uchwycenie wysokości drzewa, na poziomie którego zostaje zachwiana regularność spowodowana nagłym zwiększeniem się współczynnika. Najwcześniejszy taki moment wyznacza optymalne ucięcie dendrogramu (ang. cut-off point) i tym samym liczbę skupień. Formalnie, jest to największe takie k, dla którego spełniona jest poniższa nierówność

$$\alpha_{k+1} > \bar{\alpha} + kS_{\alpha} \tag{2.24}$$

Gdzie α_{k+1} oznacza współczynnik fuzji o numerze $k+1$, natomiast $\bar{\alpha}$ oraz S_{alpha} oznaczają odpowiednio średnią i odchylenie standardowe wartości współczynników o numerach wcześniejszyc tj. od 1 do k.

Odmiennym pomysłem, który wykorzystuje rachunek prawdopodobieństwa w odniesieniu do aglomeracyjnych metod hierarchicznych jest **indeks Dudy-Harta**, który opiera się na ilorazie:

$$DH(k) = \frac{WSS_2}{WSS_1} \tag{2.25}$$

Powyższa statystyka funkcjonuje jako **moment stopu** dla działania algorytmu. Jeśli dla danego dwupodziału wartość ilorazu jest odpowiednio niska, wówczas sprawdzany jest kolejny dwupodział zbioru, aż do momentu przekroczenia wartości krytycznej. Jest ona wyznaczana na podstawie informacji o liczebności próby, wymiarze przestrzeni oraz *p-value* dla standardowego rozkładu normalnego (por. [49, s.163]). Autorzy doszli do wniosku, że graniczną wartością jest $DH = 3,20$. Należy jednak pamiętać, że jest to jedynie **statystyka pseudotestowa**, gdyż losowy charakter wskaźników zakładany jest *ex post factum*, a nie przed procedurą tworzenia skupień.

O metodach wykorzystujących rachunek prawdopodobieństwa powiemy więcej w rozdziale poświęconym testowaniu hipotez na temat liczby klas ukrytych. Dzięki założeniu o losowym charakterze wskaźników, będzie możliwe wykorzystanie metody największej wiarogodności do oszacowania parametrów. Dodatkowo wyznaczenie wartości funkcji wiarogodności pozwoli na zastosowanie testu ilorazu wiarogodności.

Na zakończenie powiemy jeszcze o interesującej metodzie, w której liczba skupień jest traktowana jako jeden z parametrów. Idea opiera się na konstrukcji przedziałów ufności dla tego parametru. (zob. [**?**]). Procedura bazuje na metodzie **bootstrap** - wielokrotnego niezależnego doboru mniejszych prób z jedynej próby, którą dysponujemy. [7] Metoda ta należy do rodziny tzw. **testów nieparametrycznych**, które stosuje się, gdy nie znamy lub nie możemy określić rozkładu statystyki testowej. Wówczas, opierając się na założeniu, że próba jest wiernym odwzorowaniem populacji, wielokrotnie pobieramy z niej mniejsze próbki. W każdej z nich obliczamy wartość interesujących nas statystyk i na podstawie wielu doświadczeń otrzymujemy ich rozkłady.

Jako jedni z pierwszych, Peck i in. [**?**] podali przepis na budowę przedziałów ufności za pomocą opisanej poniżej metody:

1. Zdefiniuj **funkcję kosztu** opisującą wartość kryterium optymalizacyjnego (lub różnicę, między oryginalnym rozkładem, a jego rekonstrukcją)

[7] Za autora metody uznaje się Bradleya Efrona a klasycznym już dziełem jest "The jackknife, the bootstrap, and other resampling plans. Philadelphia: Pa. Society for Industrial and Applied Mathematics, 1982". Nazwa bootstrap (ang. sznurówka) pochodzi od idiomatycznego zwrotu *to pull oneself up by one's bootstraps* - wydobyć się z opresji własnymi siłami. Więcej o metodzie można przeczytać w [**?**].

2. Określ rozkład prawdopodobieństwa w próbie, na podstawie którego będą losowane mniejsze próby
3. Pobierz k niezależnych prób
4. W każdej próbie oblicz liczbę skupień, która minimalizuje wartość zdefiniowanej funkcji kosztu
5. Zbuduj empiryczny rozkład liczby uzyskanej w poprzednim kroku i wybierz $(1-\alpha)\cdot 100\%$ najbardziej prawdopodobnych wyników

Funkcja kosztu w pierwszym kroku praktycznie mieć dowolną postać. W ogólnym przypadku wygląda ona następująco:

$$L(d) = c_1(k) + \sum_{i}^{k} \int_{R_i} c_2(x, y_i)\, dP(x) \tag{2.26}$$

Pierwszy składnik funkcji $c_1(k)$ oznacza koszt z powodu liczby skupień (każdy dodatkowy parametr komplikuje model). Podobnie jak w wielu zagadnieniach mikroekonomicznych zakłada się jej liniowy charakter. Drugi składnik zawiera całkę względem miary probabilistycznej P określającej rozkład wskaźników x. Zazwyczaj ten rozkład nie jest znany [**?**, s.184] i za jego estymator przyjmuje się równomierny rozkład punktowy. Innymi słowy, zamiast całki względem miary stosuje się miarę liczącą, czyli sumę z wagami (prawdopodobieństwami) równymi $\frac{1}{n}$, gdzie n jest liczebnością próby. Każda z całek (lub równoważnie każda suma) ograniczona jest do zbioru obserwacji zaklasyfikowanych do tego skupienia skupienia (tj. $R_i = \{f(x) = i\}$). Numer skupienia oznaczony jest tu indeksem i i wyraża się liczbą naturalną ze zbioru $\{1, 2..., k\}$. Jest on przypisywany przez odpowiedni algorytm analizy skupień za pomocą funkcji $f : (x_1, x_2, ..., x_n) \to (1, 2, ..., k)$. Wektor $\mathbf{y} = (y_1, y_2, ..., y_k)$ indeksowany za pomocą numerów kolejnych skupień oznacza zbiór wartości opisujących jednostki typowe dla skupienia. W przypadku metody K-średnich jest to po prostu wektor średnich wskaźników w każdym skupieniu. Koszt c_2 jest więc tym większy, im dana obserwacja bardziej różni się od typowego reprezentanta klasy (np. od wektora średnich w klasie). Po opisanych powyżej modyfikacjach funkcja celu ma przyjmuje następującą postać:

$$L(d) = ak + \frac{1}{n} \sum_{i}^{G} \sum_{x_j \in = R_i} (x_j - y_i)^2 \tag{2.27}$$

Jak łatwo możemy zauważyć, drugi składnik prawej strony równości jest nam dobrze znaną miarą rozproszenia wewnątrzgrupowego $trWSS$. Natomiast pierwszy odgrywa rolę kompensującą spadek wartości funkcji dzięki zwiększeniu liczby skupień.

Oryginalność metody powyższej metody polega głównie na pomysłowym wykorzystaniu mocy obliczeniowej komputerów. Nie ma tu jednak żadnych założeń *a priori* odnośnie rozkładu rozkład wskaźników czy skupień. Wręcz przeciwnie, stosowana jest procedura odwrotna - probabilizacja próby w celu uzyskania rozkładu estymatorów parametrów. W dalszym ciągu bazuje się na standardowych kryteriach wariancyjnych.

2.2.4. Problem jednostek odstających. Stabilność rozwiązania.

> Od dawna zastanawiało statystyków czy rozsądnym jest odrzucenie obserwacji odstających. Nie można zaprzeczyć, że takie obserwacje istnieją, ale ich pozbycie się jest tylko jedną z wielu możliwości. ([**?**])

Definicja

Obecność jednostek odstających zawsze budziła wśród statystyków pewien niepokój i dlatego w literaturze poświęcono im wiele miejsca. Mimo, to jak podają Gallegos i David ([**?**]) samo pojęcie nie doczekało się precyzyjnej definicji. Jak podaje [**?**] *jest to taka obserwacja, która* ***wydaje się*** *znacznie odchylać od pozostałych obserwacji należących do tej samej próby.*

Nie ma też ogólnej zgody, co do terminologii. Barnett [**?**, s.243] podaje kilka odpowiedników jednostek odstających w języku angielskim: *outlier, unrepresentative, rogue, spurious, maverick*. W wielu pakietach statystycznych jednostki odstające noszą nazwę **szumu** (ang. noise) lub **zanieczyszczenia** (ang. contaminants). Wszystkie te określenia są jednoznacznie negatywne - obecność takich obserwacji jest niepożądana z punktu widzenia analiz. W tym wypadku analiza skupień nie jest wyjątkiem. Odruchowym postępowaniem, o czym powiemy więcej w dalszej części tej sekcji, byłoby zatem pozbycie się takich obserwacji ze zbioru. Jednak jak zobaczymy, nie zawsze taka decyzja jest uzasadniona.

Pochodzenie outliersów

Fenomen jednostek odstających nie jest tylko problemem statystycznym. Jest to zrozumiałe, ponieważ dotyczy ogólnego problemu dewiacji, czyli odstępstwa od normy. Trudno jednoznacznie rozstrzygnąć, które obserwacje faktycznie są odstające: klasyfikacja takich obiektów wynika z przyjętego poczucia normy i zakresu tolerancji. Przykłady realnych sporów wynikających z subiektywizmu można znaleźć we wprowadzeniu do pracy Barnetta [8].

Można wyróżnić dwa podstawowe źródła *outliersów*: przypadkowe i losowe. Do pierwszej grupy należą wszelkiego rodzaju błędy związane z ingerencją człowieka. Zalicza się tutaj błędy pomiaru lub kodowania danych (np. literówki, błędy interpunkcyjne, „czeskie błędy"). Ten przypadek nie będzie nas dalej interesował, ponieważ znajduje się poza obszarem zainteresowań statystyki.

Drugie podejście zakłada, że jednostka odstająca jest zjawiskiem losowym. Oznacza to, że jest pewną niezwykle mało prawdopodobną realizacją pewnej zmiennej losowej. Znając rozkład tej zmiennej możemy zatem wyznaczyć zakres prawdopodobnych wyników, na zasadzie podobnej do budowy przedziału ufności w estymacji przedziałowej lub obszaru przyjęcia hipotezy przy testowaniu hipotez.

Identyfikacja

Poniższy przykład pochodzi z książki [11] i ma na celu pokazanie, że zastosowanie modelu statystycznego może być niezwykle przydatne do identyfikacji jednostek odstających:

Tabela 2.12: Zbiór z „ wyraźnie" odstającą jednostką

nr	1	2	3	4	5	6	7	8	9	10
x	1,74	1,46	-0,28	-0,02	-0,4	0,02	**3,89**	1,35	-1,1	0,71

Narzucającym się kandydatem na obserwację odstającą jest numer 7. Czy jest możliwe jakiekolwiek przekonujące wyjaśnienie, czemu wybraliśmy akurat tę obserwację. Naturalnym wydaje się wyjaśnienie, że ta obserwacja jest najbardziej odległa od pozostałych. Aby nie komplikować sytuacji, przyjmijmy, że dokonaliśmy pomiaru na odpowiednio mocnej skali. Macierz odległości między obserwacjami została przedstawiona w tabeli 2.13

[8] O doniosłości tych sporów świadczy fakt, że niektóre z nich rozstrzygane były na drodze sądowej

Tabela 2.13: Macierz odległości między obserwacjami z tabeli 2.12

	1	2	3	4	5	6	7	8	9	10
1	0	.	.	.	.	.	.	.	.	.
2	0,28	0	.	.	.	.	.	.	.	.
3	2,02	1,74	0	.	.	.	.	.	.	.
4	1,76	1,48	0,26	0	.	.	.	.	.	.
5	2,14	1,86	0,12	0,38	0	.	.	.	.	.
6	1,72	1,44	0,3	0,04	0,42	0	.	.	.	.
7	2,15	2,43	4,17	3,91	4,29	3,87	0	.	.	.
8	0,39	0,11	1,63	1,37	1,75	1,33	2,54	0	.	.
9	2,84	2,56	0,82	1,08	0,7	1,12	4,99	2,45	0	.
10	1,03	0,75	0,99	0,73	1,11	0,69	3,18	0,64	1,81	0
suma	14,33	12,65	12,05	11,01	12,77	10,93	**31,53**	12,21	18,37	10,93

Nie ma wątpliwości, że średnia odległość obserwacji nr 7 od pozostałych powoduje, że powinna być ona uznana za *outliersa*. Czy zawsze jednak stosowanie odległości euklidesowej jest uzasadnione? W sekcji na temat podobieństwa obiektów pojawiła się uwaga, że daje ona miarodajne wyniki jedynie dla wybranej rodziny rozkładów. Faktycznie, jeśli założymy, że próba pochodzi z rozkładu normalnego, którego parametry są estymowane na podstawie próby), to obserwacja ta zostanie zaklasyfikowana jako odstająca.

Tabela 2.14: Wartości p-value względem dwóch rozkładów normalnych

ω_i	1	2	3	4	5	6	7	8	9	10
x	1.74	1.46	-0.28	-0.02	-0.40	0.02	**3.89**	1.35	-1.10	0.71
N(0,1)	0.041	0.072	0.390	0.492	0.345	0.492	**0.000**	0.089	0.136	0.239
$N(\hat{\mu}, \hat{\sigma})$	0.243	0.308	0.240	0.299	0.215	0.309	**0.014**	0.335	0.101	0.493

Czy jesteśmy w stanie wskazać rozkład, który zmieniłby naszą decyzję? Naturalnie, takich rozkładów jest nieskończenie wiele. Przykładem jest dowolny rozkład normalny o odpowiednio dużej wariancji. Możemy również skorzystać z pozostałych rodzin rozkładów. I tak na przykład powyższa zbiorowość jest realizacją standardowego rozkładu Cauchy'ego.[9]

Czy jednak posługiwanie się rozkładami o dużym rozproszeniu jest uzasadnione? Najczęściej nie. Uzasadnienie płynie z teorii weryfikacji hipotez. Posługiwanie się zbyt tolerancyjnym testem (o małej mocy) nie doprowadzi do nas do identyfikacji odpowiednich jednostek.

Tabela 2.15: Możliwe błędy decyzyjne przy identyfikacji jednostek odstających

		Stan faktyczny	
		$x \in out$	$x \notin out$
Decyzja	$x \in out$	OK	α
	$x \notin out$	β	OK

Dążymy zatem do tego, aby móc posługiwać się regułą decyzyjną Neymana-Pearsona do weryfikacji hipotez na temat posiadania przez każdą z obserwacji cech jednostki odstającej. Aby móc skonstruować odpowiednie funkcje decyzyjne, niezbędne jest założenie o losowym

[9] Gęstość rozkładu Cauchy'ego wyraża się wzorem: $f(x) = \frac{1}{\pi(x^2+1)}$

charakterze jednostek odstających. Dopiero na tej podstawie możemy próbować stworzyć odpowiedni model statystyczny.

Ilustrację takiego podejścia można znaleźć w [**?**]. Zarówno idea, jak i terminologia przypomina klasyczny model weryfikacji hipotez. Na początku definiowany jest odpowiednik obszaru krytycznego - tzw. **obszar odstający** (ang. outlier region).

2.2.1. Definicja. Dwustronnym obszarem odstającym na poziomie α dla rozkładu zmiennej losowej X o dystrybuancie F, wartości oczekiwanej μ i wariancji σ^2 nazywamy zbiór:

$$out(\alpha, \mu, \sigma) = \{x : \frac{|x-\mu|}{\sigma} > F_X^{-1}\left(1-\frac{\alpha}{2}\right)\} \tag{2.28}$$

W przypadku, gdy X ma rozkład normalny, prawa strona nierówności opisana jest przez odpowiedni kwantyl standardowego rozkładu normalnego. W tym przypadku hipoteza zerowa głosi, że dana obserwacja należy do obszaru przyjęcia:

$$in(\alpha, \mu, \sigma) = \{x : \frac{|x-\mu|}{\sigma} \leqslant F_X^{-1}\left(1-\frac{\alpha}{2}\right)\} \tag{2.29}$$

W ten sposób problem identyfikacji jednostek odstających sprowadziliśmy do wyznaczenia dwóch liczb: górnego i dolnego krańca pewnego przedziału w oparciu n-elementową próbę i ustalony poziom istotności α. Wielkość $1-\alpha$ wyznacza nam prawdopodobieństwo, że w zbiorze nie ma żadnych outlierów.

Jądrowa estymacja gęstości

Zastosowanie metod statystycznych do problemu wykrywania jednostek odstających może okazać się skuteczne już na poziomie niezbyt wyrafinowanych sytuacji. Dla przykładu, rozpatrzmy następujący zbiór:

Tabela 2.16: Zbiór z „ukrytą" jednostka odstającą

nr	1	2	3	4	5	6	7	8	9
x	1,1	1,5	1,76	2	4	6	6,27	6,9	7,1

Macierz odległości dla tego zbioru przedstawia się następująco:

Tabela 2.17: Macierz odległości euklidesowych między obserwacjami

	1	2	3	4	5	6	7	8	9
1	0	.	.	.	.	.	.	.	.
2	0,5	0	.	.	.	.	.	.	.
3	0,4	0,9	0	.	.	.	.	.	.
4	0,26	0,24	0,66	0	.	.	.	.	.
5	2,5	2	2,9	2,24	0	.	.	.	.
6	4,77	4,27	5,17	4,51	2,27	0	.	.	.
7	5,6	5,1	6	5,34	3,1	0,83	0	.	.
8	4,5	4	4,9	4,24	2	0,27	1,1	0	.
9	5,4	4,9	5,8	5,14	2,9	0,63	0,2	0,9	0
suma	23,93	21,91	26,73	22,63	**19,91**	22,72	27,27	21,91	25,87

Zgodnie ze wcześniejszym rozumowaniem „euklidesowym" najmniej prawdopodobnym kandydatem na jednostkę odstającą jest obserwacja nr 5. Średnia odległość od pozostałych jednostek jest najmniejsza, ponieważ wartość 4 jest medianą a zatem minimalizuje sumę odległości od reszty obserwacji. Gdy jednak przeanalizujemy histogram, zobaczymy, że leży ona w obszarze o najmniejszej gęstości prawdopodobieństwa. Innymi słowy znajduje się „z dala" od pewnych dwóch „skupisk" obserwacji.

Przybliżenie histogramu ilustruje nam jak wygląda hipotetyczna krzywa gęstości prawdopodobieństwa dla uzyskanej próby. Taka procedura „wygładzania" nazywana jest **jądrową estymacją gęstości**(ang. kernel density estimation). Przystępne wprowadzenie do tej metody można znaleźć m.in w [24] lub [32].

Intuicja, która stoi za tą metodą opiera się o założenie, że populacja nie jest jednorodna. Oznacza to, że rozkład zmiennych w próbie jest efektem mieszania się wielu składników, a podstawowym zadaniem jest dotarcie do generatorów lub jąder (stąd ang. kernel) tych elementarnych rozkładów.

W kontekście analizy jednostek odstających poszukuje się więć nie jednego, ale pewnej liczby rozkładów. Dzięki nim wyznaczane są obszary o dużej gęstości oraz obszary odstające. Równanie takiego modelu ma prostą postać:

$$f(x) := \lambda F + (1 - \lambda)G \qquad , \lambda > 0.5 \tag{2.30}$$

Gdzie F jest gęstością rozkładu obserwacji „normalnych", natomiast G - odstających. Powyższy model dość łatwo można uogólnić na skończoną liczbę rozkładów. Co więcej, możemy opuścić ograniczenie na współczynniki proporcji λ. W ten sposób doszliśmy do dobrze nam znanego wyniki z sekcji poświęconej identyfikowaniu struktury w zbiorze:

$$f_1(x) := \sum_{j}^{j} \lambda_j \cdot f_j(x) \tag{2.31}$$

Związek ten nie jest przypadkowy. Model analizy skupień opisany za pomocą kombinacji rozkładów jest w rzeczywistości uogólnieniem metody włączenia outlierów do analizy poprzez wskazanie częstości ich występowania i rozkładu, który je opisuje.

Jak zatem skutecznie identyfikować jednostki odstające? Prawdopodobnie podejście statystyczne nie jest jedyną słuszną możliwością, ale z pewnością proponuje jasne uzasadnienia dla podejmowanych decyzji.

Na koniec przedstawimy jeszcze jedną metodę rozpoznawania takich jednostek. Do tego wykorzystamy wprowadzone wcześniej kryterium **indeksu sylwetki**. Zdefiniujmy jednostkę odstającą jako obserwację, która „nie pasuje" do żadnego ze skupień. Oznacza to, że bez względu na to, do jakiej klasy ją przypiszemy popełnimy błąd objawiający się spadkiem kryterium optymalizującego. W naszym przypadku tym kryterium jest właśnie wartość indeksu sylwetki. Jak pamiętamy, wartości bliskie zera oznaczają, że jednostka znajduje się „na pograniczu" dwóch skupień, natomiast ujemne wartości wskazują na błędne przyporządkowanie danej obserwacji. Dla przykładu obejrzyjmy wyniki dla klasyfikacji rozmytej metodą **Fanny**:

Widzimy, że dla 19 jednostek indeks sylwetki nie przekroczył **0.3**, z czego dla dwóch pokazał ujemną wartość. Naturalnie, na skutek losowego charaktery outlierów, część z nich „znalazła się" w wyróżnionych skupieniach, stąd wysokie wartości współczynników dla pozostałych obserwacji.

Tabela 2.18: Fanny: Klasyfikacja rozmyta jednostek odstających

	numer skupienia									
nr obserwacji	1	2	3	4	5	6	7	8	9	indeks sylwetki
270	0.03	0.04	0.04	0.06	0.06	0.08	0.09	0.14	0.44	0.495
271	0.05	0.07	0.06	0.09	0.06	0.1	0.11	0.29	0.16	0.5
272	0.04	0.05	0.05	0.06	0.05	0.07	0.11	0.41	0.16	0.613
273	0.04	0.06	0.05	0.13	0.27	0.23	0.06	0.08	0.09	0.148
274	0.05	0.07	0.05	0.4	0.13	0.14	0.05	0.06	0.05	0.627
275	0.05	0.06	0.06	0.09	0.09	0.13	0.1	0.14	0.27	0.445
276	0.07	0.12	0.09	0.12	0.07	0.1	0.11	0.2	0.11	0.365
277	0.11	0.35	0.22	0.05	0.04	0.04	0.06	0.07	0.05	0.229
278	0.06	0.07	0.07	0.11	0.12	0.14	0.11	0.13	0.19	0.274
279	0.12	0.3	0.13	0.1	0.06	0.07	0.07	0.08	0.06	0.534
280	0.01	0.02	0.01	0.06	0.74	0.09	0.02	0.02	0.02	0.6320
281	0.22	0.14	0.22	0.06	0.05	0.06	0.1	0.09	0.07	-0.048
282	0.05	0.07	0.06	0.11	0.13	0.17	0.1	0.13	0.18	0.115
283	0.05	0.06	0.05	0.13	0.23	0.23	0.07	0.09	0.1	0.021
284	0.05	0.06	0.06	0.11	0.12	0.17	0.1	0.13	0.2	0.135
285	0.18	0.13	0.21	0.06	0.05	0.06	0.12	0.1	0.08	0.118
286	0.12	0.11	0.15	0.07	0.06	0.07	0.17	0.13	0.11	0.183
287	0.05	0.06	0.05	0.12	0.11	0.21	0.09	0.13	0.17	0.152
288	0.11	0.1	0.14	0.07	0.05	0.06	0.23	0.14	0.11	0.375
289	0.21	0.3	0.17	0.06	0.04	0.05	0.06	0.06	0.05	0.202
290	0.17	0.18	0.48	0.03	0.02	0.02	0.04	0.04	0.03	0.469
291	0.03	0.05	0.04	0.33	0.12	0.28	0.04	0.06	0.06	0.171
292	0.05	0.05	0.06	0.05	0.04	0.05	0.41	0.15	0.14	0.453
293	0.11	0.22	0.12	0.14	0.08	0.09	0.07	0.09	0.07	0.291
294	0.06	0.07	0.07	0.11	0.13	0.15	0.11	0.12	0.16	0.105
295	0.04	0.05	0.06	0.04	0.03	0.04	0.53	0.12	0.09	0.586
296	0.12	0.11	0.14	0.07	0.06	0.07	0.19	0.13	0.11	0.265
297	0.12	0.12	0.18	0.06	0.05	0.06	0.18	0.14	0.09	-0.027
298	0.03	0.04	0.04	0.15	0.11	0.45	0.05	0.07	0.07	0.519
299	0.1	0.21	0.19	0.07	0.05	0.06	0.11	0.13	0.08	0.081
300	0.05	0.06	0.06	0.13	0.21	0.22	0.07	0.09	0.11	0.011

Sposoby postępowania z jednostkami odstającymi

Identyfikacja jednostek odstających to jedynie pierwsza część procesu decyzyjnego. Kolejnym etapem jest decyzja odnośnie ich dalszej obecność w zbiorze. Nie ma badania, w którym nie istniały by ekstremalne przypadki, ale ich eliminacja jest tylko jednym z całej gamy dostępnych rozwiązań. Nie podajemy tutaj uniwersalnej recepty lecz opisujemy krótko wady i zalety każdej z możliwych dróg postępowania. Można wyróżnić dwa najczęstsze sposoby radzenia sobie z odstającymi przypadkami. Oto one:

1. Eliminacja. Z punktu widzenia analizy skupień manipulacja jednostkami odstającymi powinna odbywać się ze szczególną ostrożnością. Czasami pochopna eliminacja outlierów może oznaczać pozbycie się istotnej klasy obserwacji. Z kolei ignorowanie takich jednostek może powodować zaburzenia w działaniu niektórych algorytmów.

2. Włączenie. Traktujemy outliera jako pełnoprawną obserwację. Inkluzja oznacza poszukiwanie takiego rozkładu, który gwarantowałby homogeniczność dla całej próby bądź danej klasy. Innymi słowy, staramy się tak manipulować parametrami obszaru odstającego, aby zminimalizować prawdopodobieństwo należenia do niego dla jak największej liczby obserwacji.

3. Adaptacja. Rezygnujemy ze stosowania wrażliwych miar (np. średniej) i redukujemy informację na temat położenia obiektów w przestrzeni do wzajemnych relacji między nimi. Innymi słowy, przestają nas interesować dokładne współrzędne obserwacji, ale skupiamy się na ich uporządkowaniu w przestrzeni. W wyniku takiej operacji może zdarzyć się, że obserwacje które były odległe od siebie „nominalnie" o jednostkę będą traktowane podobnie jak obserwacje odległe od siebie o sto jednostek. W konsekwencji posługujemy się innymi parametrami rozkładu np. medianą.

Jednostki odstające a klasyczne metody Można powiedzieć, że klasyczne metody zaprojektowane są z myślą o analizowaniu zbiorów pozbawionych jednostek odstających. Oznacza to, że nie wykorzystywane są żadne procedury filtrujące właściwe obserwacje. Mamy więc prawo podejrzewać, że końcowy rezultat obarczony jest pewnym błędem. Na szczęście w niektórych przypadkach jesteśmy w stanie odtworzyć proces formowania się skupień i zidentyfikować potencjalne obserwacje odstające.

Metody hierarchiczne Tak jest na przykład w przypadku metod hierarchicznych. Pożyteczna jest analiza końcowego rezultatu - dendrogramu, a dokładniej kolejnych współczynników fuzji. Dla każdego obiektu można wyznaczyć liczbę połączeń jakie odnotowały kolejne skupienia, w których się on znajdował z obiektami w pozostałych skupieniach. Jeśli ta liczba jest bliska jedności, oznacza to, że obiekt został włączony dopiero w ostatnich fazach algorytmu.

Niestety, nie każda metoda hierarchiczna prowadzi do jasnej identyfikacji. Metoda Warda, średniego wiązania i za pomocą centroidów zwykle nie zdają egzaminu z identyfikacji outlierów. W obu przypadkach posługują się one zaburzonymi odległościami. Oznacza to, że w przypadku braku jednostek odstających rozpoznana struktura zbioru byłaby zupełnie inna. Poniżej możemy zauważyć przykład takiego zaburzenia.

Natomiast jeśli chodzi o metodę pojedynczego wiązania, to paradoksalnie, to co jest jej największą przypadłością - tworzenie sztucznych łańcuchów, w przypadku identyfikacji outlierów okazuje się niezwykle użyteczne. Tendencja do wyznaczania jednego dużego spójnego podzbioru i pozostałych wyraźnie mniej licznych skupień.

Metoda K-średnich Najmniej odporna na obecność jednostek odstających wydaje się metoda K-średnich. O pierwszej przyczynie zdążyliśmy już powiedzieć. Jak wiadomo, algorytm dąży do minimalizacji kwadratów wewnątrzgrupowych (patrz: metody wariancyjne), w związku z czym posługuje się miarą wrażliwą na pojedyncze odchylenia.

Druga przyczyna jest immanentną cechą algorytmu. Jest nią generalny brak stabilności rozwiązań względem warunków początkowych. Konsekwencją tej właściwości jest jedynie lokalna optymalność rozwiązań. Ta własność przysługuje algorytmowi bez względu na obecność czy brak jednostek odstających. Niefortunny wybór punktów startowych rodzi niebezpieczeństwo, że jednym z nich będzie właśnie obserwacja odstająca, wobec czego już w pierwszym kroku powstaje niepożądane "zakotwiczenie". To ryzyko jest proporcjonalne do częstości takich jednostek w zbiorze.

2.3. Wnioski

Krótkie podsumowanie przeglądu rozwiązań fundamentalnych kwestii analizy skupień zaczniemy od przedstawienia tabeli 2.18. Ujmuje ona w syntetyczny sposób reakcję klasycznych metod na podstawowe postulaty.

Tabela 2.19: Realizacja postulatów dobrego podziału przez wybrane metody analizy skupień

typ metod	Kryteria dobrego podziału			
	segmentowalność	optymalność	liczba skupień	jednostki odstające
hierarchiczne	**brak bezpośrednich kryteriów**	brak bezpośrednich kryteriów, współczynnik korelacji kofenetycznej	ustalana apriori, kryterium stepsize, indeks Mojeny	**brak bezpośrednich kryteriów**
k-średnich	**brak bezpośrednich kryteriów**	suma kwadratów odległości wewnątrzgrupowych (WSS), indeks sylwetki	ustalana apriori, wykres "osypiska", statystyka odstępu, współczynniki oparte na porównaniu parami WSS dla kolejnych skupień	**brak bezpośrednich kryteriów**
modele blokowe	współczynnik krystalizacji struktury	współczynnik krystalizacji struktury	współczynnik krystalizacji struktury	**brak bezpośrednich kryteriów**
możliwe usprawnienia:	wprowadzenie modelu statystycznego i sformułowanie problemu w języku testowania hipotez na temat braku struktury	współczynnik separowalności wiążący BSS i WSS, point-biserial correlation	statystyki pseudotestowe F (np. indeks Dudy-Harta), przedziały ufności	wprowadzenie modelu statystycznego i sformułowanie problemu w języku testowania hipotez na temat istnienia takich jednostek

Rozdział 3

Probabilistyczny model analizy skupień

> Warto byłoby wprowadzić system analizy skupień bez arbitralnych założeń na temat podobieństwa obiektów - Wolfe [62, s.329]

Banfield i Raftery [10] sygnalizują, że przez długi okres analiza skupień rozwijała się głównie poprzez nowe pomysły i empiryczne badania metod *ad hoc*, z dala od bardziej formalnych procedur statystycznych. W ciągu ostatnich kilku lat odkryto, że oparcie analizy skupień na modelu probabilistycznym może być użyteczne zarówno dla ewaluacji istniejących metod, jak również może służyć jako bodziec do wprowadzenia metod zupełnie nowych.

W poprzednim rozdziale wykazaliśmy potrzebę założenia probabilistycznego związku między wskaźnikami a cechą ukrytą. Takie podejście nie gwarantuje jednoznacznej odpowiedzi na wszystkie problemy analizy skupień, jednak na pewno stanowi przełom w formułowaniu uzasadnień dla niektórych ważnych decyzji. O tych sytuacjach mówiliśmy w poprzednim rozdziale. Pokazaliśmy, że zastosowanie modeli probabilistycznych może być użyteczne przy weryfikacji hipotez na temat braku struktury w zbiorze oraz identyfikacji jednostek odstających. Bazowaliśmy wówczas na założeniu, że zaobserwowana próba jest realizacją pewnej kombinacji zmiennych losowych o różnych rozkładach. W dalszej części przedyskutujemy przydatność tych metod przy kolejnym, ważnym zagadnieniu wyznaczania optymalnej liczby skupień.

3.1. Parametryzacja modelu

Idea, która stoi za stochastycznym modelem analizy skupień jest rozwinięciem koncepcji modelowego wykrywania struktury w zbiorze lub istnienia jednostek odstających. Podobnie, podstawowym założeniem jest fakt, że łączny rozkład wskaźników w próbie jest kombinacją pewnych niezależnych rozkładów. Każdy taki rozkład odpowiada konkretnemu poziomowi cechy ukrytej. Innymi słowy, zakładamy, że próba składa się z obserwacji, które pochodzą z pewnych nieobserwowalnych klas. Mamy więc do czynienia z mieszaniną dwóch populacji, w której *zgubiliśmy* dane na temat pochodzenia danej obserwacji i naszym zadaniem jest wyodrębnić jej elementy (ang. unmix the mixture, [6]).

Zadanie polega na identyfikacji składników mieszaniny, czyli wyznaczeniu liczby i proporcji wartości cechy ukrytej oraz jej parametrów, które wyznaczają łączny rozkład wskaźników. Dla określenia łącznego rozkładu wskaźników będziemy mówić o **profilu** obserwacji. Formalnie, model opisuje się następująco:

$$f(\mathbf{x}) = \sum_{j=1}^{k} \lambda_j \cdot f(\mathbf{x}|\theta_\mathbf{j}) \tag{3.1}$$

Z punktu widzenia problemu analizy skupień, interesować nas będzie nie samo prawdopodobieństwo uzyskania konkretnego profilu przy danym poziomie cechy ukrytej, ale prawdopodobieństwo przynależności do danej klasy ukrytej pod warunkiem uzyskania konkretnego profilu: $f(j|\mathbf{x})$. Korzystając z reguły Bayesa otrzymujemy zależność:

$$f(j|\mathbf{x}) = \frac{f(\mathbf{x}|\theta_\mathbf{j})\lambda_\mathbf{j}}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\theta_\mathbf{j})\lambda_\mathbf{j}}{\sum_{j=1}^{k} \lambda_j f(\mathbf{x}|\theta_\mathbf{j})} \tag{3.2}$$

Wyliczenie odpowiednich warunkowych prawdopodobieństw jest stosunkowo proste przy ustalonych parametrach modelu tj. gdy dysponujemy rozkładem brzegowym cechy ukrytej oraz wektorem warunkowych prawdopodobieństw dla każdej jej wartośći. Przypomnijmy jednak, że w analizie skupień jedyną dostępną informacją jest wiedza na temat łącznego rozkładu wskaźników.

Niezależnie od charakteru modelu najbardziej interesować nas będą jego cztery komponenty:

1. Rozkład łączny wskaźników w całej populacji
2. Brzegowy rozkład cechy ukrytej (równoważnie proporcje klas ukrytych w populacji)
3. Warunkowe prawdopodobieństwo łącznego rozkładu wskaźników w każdej klasie ukrytej
4. Prawdopodobieństwo przynależności profilu do danej klasy ukrytej

Warto podkreślić, że ten ostatni element z punktu widzenia klasycznej analizy jest tylko produktem ubocznym niezbędnym do estymacji, natomiast w analizie skupień jest czymś, czego w istocie szukamy.

3.1.1. Model dyskretny

Postać modelu

Zacznijmy od wprowadzenia następujących oznaczeń. Dla ustalenia uwagi wyobraźmy sobie, że mamy trzy zmienne wskaźnikowe: A, B, C o odpowiednio I, J, K wartościach oraz jedną ukrytą zmienną X o T wartościach, która jednoznacznie wyznacza warunkowe rozkłady wskaźników. Za każdym razem prawdopodobieństwo będziemy oznaczać grecką literą π natomiast częstość występowania w próbie oznaczać będziemy przez p. W obu przypadkach będziemy stosowali następujące indeksowanie:

$$\pi_{ijkt}^{ABCX} \equiv P(A = i, B = j, C = k, X = t)$$
$$\pi_{it}^{\bar{A}X} \equiv P(A = i|X = t)$$

Odpowiednikiem profilu obserwacji $\mathbf{x}$ będzie napis złożony z wartości ijk co implikuje, że $f(\mathbf{x})$ będzie zastąpione przez π_{ijk}. Współczynniki λ_j będą wyrażone przez rozkład brzegowy zmiennej ukrytej π_t^X

W ten sposób możemy przepisać ogólne równanie modelu jako:

$$\begin{aligned}\pi_{ijk} &= \sum_{t=1}^{T} \pi_t^X \pi_{ijk}^{\bar{A}\bar{B}\bar{C}X} \\ &= \sum_{t=1}^{T} \pi_t^X \pi_{it}^{\bar{A}X} \pi_{jt}^{\bar{B}X} \pi_{kt}^{\bar{C}X}\end{aligned}$$

Równoważność zapisów wynika bezpośrednio z aksjomatu lokalnej niezależności wskaźników. Suma prawdopodobieństw brzegowych zmiennej ukrytej oraz suma prawdopodobieństw warunkowych wskaźników prawdopodobieństwo przynależności profilu do danej klasy ukrytej sumuje się do jedności w każdej klasie ukrytej:

$$\begin{aligned}\sum_{t=1}^{T} \pi_t^X &= 1 \\ \sum_{i=1}^{I} \pi_{it}^{\bar{A}X} &= 1 \\ \sum_{j=1}^{J} \pi_{jt}^{\bar{B}X} &= 1 \\ \sum_{k=1}^{K} \pi_{it}^{\bar{C}X} &= 1\end{aligned}$$

Bez straty ogólności możemy założyć, ze wszystkie prawdopodobieństwa są dodatnie. Pozwoli nam to zastosować regułę Bayesa dla wyznaczania prawdopodobieństwa przynależności do danej klasy ukrytej t pod warunkiem uzyskania profilu ijk, co możemy zapisać jako:

$$\pi_{ijkt}^{ABC\bar{X}} = \frac{\pi_{ijkt}^{ABCX}}{\pi_{ijkt}}$$

3.1.2. Model ciągły

Modele ciągłe, w przeciwieństwie do przedstawionych powyżej modeli dyskretnych, operują pojęciem gęstości prawdopodobieństwa. Gęstość profilu $\mathbf{x}$ będziemy zapisywać jako $f(\mathbf{x})$. Model zakłada, że łączna gęstość jest wypukłą kombinacją funkcji gęstości pewnych rozkładów. Rozkłady te mogą (choć nie muszą) należeć do tej samej rodziny (np. wykładniczej). Wszystkie muszą dać się opisać skończoną liczbą parametrów. W praktyce, do modelowania wykorzystuje się różne rozkłady. Z uwagi na pewne pożądane własności i łatwość opisu, największą popularnością cieszy się rodzina rozkładów normalnych.

Podstawowe klasy modeli zostały uporządkowane i opisane w pracy [10] a ich opis został rozwinięty w [25]. Jedynym problematycznym podzbiorem parametrów modelu są macierze kowariancji. W porównaniu z wektorem średnich są one wielowymiarowe, co powoduje, że do ich specyfikacji wymagana jest również informacja na temat współzależności wektorów. Pomysł zaproponowany w cytowanych powyżej pracach opiera się o dekompozycję odwracalnej macierzy kowariancji Σ:

$$\Sigma = \lambda D A D^{-1} \tag{3.3}$$

Przy czym macierz D jest ortogonalna, zatem powyższy wzór jest równoważny:

$$\Sigma = \lambda D A D^T \tag{3.4}$$

W obu równania λ jest pewną stałą i służy do przeskalowania iloczynu macierzy. Odpowiada on za **rozmiar** skupienia. Macierz D składa się z wektorów własnych, a diagonalna macierz A z odpowiadających im wartości własnych. Współrzędne wektorów zapisanych w D wyznaczają **orientację** głównych składowych macierzy Σ a elementy macierzy A wyznaczają **kształt** krzywej gęstości opisujących dane skupienie. Za pomocą tych trzech składników: orientacji, kształtu i wielkości można zdefiniować klasy różnych modeli. Krótkie podsumowanie zawiera poniższa tabela, por. [25, s.581]:

Tabela 3.1: Klasyfikacja modeli będących kombinacjami rozkładów normalnych. Źródło: [25] oraz pomoc do pakietu MCLUST w środowisku R

nazwa modelu (kod w postaci: λ, D, A)	Σ	Kształt	Orientacja	Rozmiar
EII	λI	-	-	jednakowy
VII	$\lambda_j I$	-	-	zmienny
EEI	$\lambda D I D^T$	-	jednakowa	jednakowy
VEI	$\lambda_j D I D^T$	-	jednakowa	zmienny
EVI	$\lambda D_j I D_j^T$	-	zmienna	jednakowy
VVI	$\lambda_j D_j I D_j^T$	-	zmienna	zmienny
EEE	$\lambda D A D^T$	jednakowy	jednakowa	jednakowy
EEV	$\lambda D A_j D^T$	zmienny	jednakowa	jednakowy
VEV	$\lambda_j D A_j D^T$	zmienny	jednakowa	zmienny
VVV	$\lambda_j D_j A_j D_j^T$	zmienny	zmienny	zmienny

Jeśli posłużymy się rodziną rozkładów normalnych do ilustracji zagadnienia modelowania analizy skupień za pomocą mieszaniny rozkładów normalnych to otrzymamy ogólne modelu:

$$f(\mathbf{x}) = \sum_{j=1}^{k} \lambda_j f(\mathbf{x}|\mu_j, \Sigma_j) \tag{3.5}$$

W tym przypadku μ_j jest p-wymiarowym wektorem wartości oczekiwanych, a Σ_j macierzą wariancji-kowariancji między p - wskaźnikami, jeśli obserwacja pochodzi z segmentu o numerze j. Przypomnijmy że, gęstość p-wymiarowego rozkładu normalnego opisana jest wzorem:

$$f(x_1, x_2, ..., x_p) = (2\pi)^{-\frac{p}{2}} \cdot (\det \Sigma)^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Podobnie jak w modelu dyskretnym, bez straty ogólności możemy założyć, ze wszystkie gęstości są dodatnie. Pozwoli nam to zastosować regułę Bayesa dla wyznaczania prawdopodobieństwa przynależności do danej klasy ukrytej j pod warunkiem uzyskania profilu $\mathbf{x}$, co możemy zapisać jako:

$$P(j|\mathbf{x}) = \frac{\lambda_j f(\mathbf{x}|\theta_j)}{\sum_{j=1}^{G} \lambda_j f(\mathbf{x}|\mathbf{mu}_j, \Sigma_j)} = \frac{\lambda_j f(\mathbf{x}|\theta_j)}{f(\mathbf{x})} \tag{3.6}$$

3.2. Identyfikowalność modelu

Wśród podstawowych problemów modelowania parametrycznego ważne miejsce zajmuje kwestia jednoznaczności oszacowanych parametrów. Powstaje pytanie na temat ich **identyfikowalności**. Jeśli nie są one identyfikowalne, pojawia się istotny problem wyboru najlepszego rozwiązania spośród wszystkich dających identyczny wynik. Użyteczne będzie wprowadzenie następujących definicji (por. [65] lub [61]).

3.2.1. Definicja. Model $\{X, P_\theta : \theta \in \Theta\}$ nazywamy **identyfikowalnym**, gdy istnieje jednoznaczne odwzorowanie przestrzeni parametrów w rodzinę rozkładów tj. $\forall \theta_1, \theta_2 \in \Theta$ nierówność $\theta_1 \neq \theta_2$ implikuje $P_{\theta_1} \neq P_{\theta_2}$

P_θ jest pewną rodziną rozkładów indeksowaną za pomocą wektora parametrów θ. Innymi słowy, powiemy, że model jest identyfikowalny, jeśli dany rozkład nie może być generowany przez różne zbiory parametrów.

Zagadnienie jednoznaczności rozwiązania w kontekście analizy skupień różni się w zależności od wybranego modelu. Dlatego oddzielnie zajmiemy się analizą klasy ukrytej (model dyskretny) i skończonymi mieszaninami rozkładów (model ciągły).

3.2.1. Model dyskretny

Wróćmy na moment do pierwszego rozdziału, w którym wprowadziliśmy koncepcję modelu klasy ukrytej i opisaliśmy ją za pomocą układów równań. Doszliśmy do wniosku, że dla s wskaźników zero-jedynkowych rozwiązanie zagadnienia sprowadza się do problemu wyznaczenia $(q + qs) = q(1 + s)$ parametrów: q liczby klas ukrytych oraz qs prawdopodobieństw warunkowych dla każdej z q klas.

Na podstawie odpowiednich twierdzeń z algebry liniowej parametry modelu można jednoznacznie zidentyfikować, gdy liczba niewiadomych nie przekracza liczby równań tzn. gdy spełniona jest zależność:

$$q(1 + s) \leqslant 2^s \tag{3.7}$$

Model ze wskaźnikami dychotomicznymi został uogólniony w pracy Goodmana w 1974 roku [28] na zmienne o dowolnej skończonej liczbie poziomów wartości wskaźników i cechy ukrytej. Dla ustalenia uwagi opisuje on model z czterema zmiennymi obserwowalnymi: A, B, C, D i jedną klasą ukrytą X które liczą sobie odpowiednio I, J, K, L i T poziomów wartości. Rozkład każdej ze zmiennych może być zatem opisany za pomocą $I-1, J-1, K-1, L-1, T-1$ wartości. Zatem do wyznaczenia modelu klasy ukrytej potrzebujemy $T - 1 + (I + J + K + L - 4)T = (I + J + K + L - 3)T - 1$ parametrów. Są to tzw. **parametry bazowe**. Z drugiej strony rozkład łączny wskaźników bez uwzględnienia modelu klasy ukrytej definiuje się za pomocą $IJKL - 1$ parametrów. W przypadku, gdy $IJKL < (I + J + K + L - 3)T$ mamy do czynienia ze „zbyt dużym" zbiorem parametrów model i jesteśmy w stanie znaleźć dwie różne parametryzacje równie dobrze przybliżające łączny rozkład wskaźników. Gdy $IJKL = (I + J + K + L - 3)T$ istnieje dokładnie jedno rozwiązanie problemu doboru parametrów, dzięki któremu otrzymujemy dokładne oszacowanie. Najbardziej interesujący jest jednak przypadek, gdy $IJKL > (I+J+K+L-3)T$ a właściwie $IJKL >> (I+J+K+L-3)T$ tj. gdy liczba parametrów jest wyraźnie niższa od liczby wszystkich możliwych wyników doświadczenia. Wtedy bowiem mamy do czynienia z problemem modelowania i wyboru odpowiednich estymatorów, które będą spełniać wcześniejszą funkcję parametrów bazowych.

Ponieważ estymatory są zmiennymi losowymi to przybliżenie uzyskane, za ich pośrednictwem jest zawsze obarczone pewnym losowym błędem. Fakt, że do wyznaczenia rozkładu

łącznego posługiwać się teraz będziemy czymś „bardziej zróżnicowanym" niż ustalony parametr doprowadził do powstania użytecznego pojęcia lokalnej identyfikowalności estymatorów.

3.2.2. Definicja. Element θ_i wektora parametrów θ jest **lokalnie identyfikowalny** jeśli istnieje jego otwarte otoczenie, w którym nie istnieje taki θ_j, $j \neq i$, że $P_{\theta_i} = P_{\theta_j}$.

Podamy teraz warunek lokalnej identyfikowalności. Wprowadźmy następującą konwencję. Niech π_{ijkl} oznacza rodzinę funkcji, której argumentami są parametry bazowe: $(\pi_t^X, \pi_{it}^{\bar{A}X}, \pi_{jt}^{\bar{B}X}, \pi_{kt}^{\bar{C}X}, \pi_{lt}^{\bar{D}X})$. Jest ich dokładnie $IJKL$. Warunek sprawdzany jest poprzez badanie macierzy pochodnych cząstkowych (macierzy Jacobiego) względem parametrów bazowych. Taka macierz składa się z $IJKL-1$ wierszy i $(I+J+K+L-3)T-1$ kolumn. Dla przykładu z czterema wskaźnikami dychotomicznymi i dwiema klasami ukrytymi macierz ta jest wymiaru 15×9. Element (1,1) takiej macierzy jest opisany następująco:

$$\frac{\partial \pi_{1111}}{\partial \pi_1^X} = \frac{\partial \sum_{t=1}^{2} \pi_1^X \pi_{1t}^{\bar{A}X} \pi_{1t}^{\bar{B}X} \pi_{1t}^{\bar{C}X} \pi_{1t}^{\bar{D}X}}{\partial \pi_1^X} = \quad = \pi_{11}^{\bar{A}X} \pi_{11}^{\bar{B}X} \pi_{11}^{\bar{C}X} \pi_{11}^{\bar{D}X} - \pi_{12}^{\bar{A}X} \pi_{12}^{\bar{B}X} \pi_{12}^{\bar{C}X} \pi_{12}^{\bar{D}X} \tag{3.8}$$

Jak podaje Goodman, jeśli rząd tej macierzy jest równy liczbie kolumn, wówczas estymatory największej wiarogodności parametrów bazowych są lokalnie identyfikowalne. Do testowania danego zestawu parametrów wykorzystuje się wówczas statystykę X^2 Pearsona:

$$X^2 = 2 \sum_{ijkl} np_{ijkl} \cdot \log \frac{np_{ijkl}}{n\hat{\pi}_{ijkl}} \tag{3.9}$$

Jej asymptotyczny rozkład pod warunkiem hipotezy zerowej jest rozkładem chi-kwadrat o liczbie stopni swobody równej różnicy między liczbą wierszy i liczbą kolumn powyżej opisanej macierzy, czyli $IJK - (I+J+K+L-3)T$. Zauważmy, że gdy macierz ta jest kwadratowa, wówczas liczba stopni swobody wynosi zero, co oznacza, że nie mamy do czynienia z sytuacją modelowania.

3.2.2. Model ciągły

Identyfikowalność dla modelowania za pomocą kombinacji gęstości wygląda nieco mniej przejrzyście niż w przypadku dyskretnym. Jednym z pierwszych statystyków, który podjął się teoretycznego wyjaśnienia problemu był Teicher [60], który zauważył, że *w ogólnym przypadku* mieszanina za pomocą rodziny rozkładów Gamma (w tym normalnych i dwumianowych) jest nieidentyfikowalna. Jednocześnie zaznaczył, że przy pewnych warunkach (addytywność składników i jednowymiarowość przestrzeni parametrów) jednoznaczność ta jest zagwarantowana (zob. [59]).

Jako rozwinięcie poprzednich twierdzeń Teicher wykazał, że możliwa jest jednoznaczność dla *skończonej* mieszaniny rozkładów i to dla szerokiej klasy rozkładów. Koncepcja ta została uogólniona przez Bruniego i Kocha [15], którzy rozszerzyli ją do przypadku wielowymiarowego wektora paramterów oraz wykazali, że dla pewnych topologii wykonalna jest identyfikacja dla rozkładów spoza rodziny normalnej.

Z kolei Arminger i Kusters [5, s.383] zgadzają się z pierwszą tezą Teichera, twierdząc, że w ogólnym przypadku modelu klasy ukrytej nieznane są kryteria dla identyfikacji wszystkich parametrów. Zalecają stosowanie metody pokrewnej, która zaproponował Goodman tj. aby badać strukturę tzw. **macierzy informacyjnej** Fishera. W statystyce, informacja Fishera odgrywa ważną rolę. Jest ona jedną z podstawowych miar ilości informacji jaką niesie rozkład zmiennej na temat parametru, od którego zależy wartość funkcji wiarogodności. Informacja dla pojedynczej obserwacji wyraża się poprzez:

$$I(\theta) = D^2(\frac{\partial}{\partial\theta}\log L(\theta, \mathbf{x})) = E(\frac{\partial}{\partial\theta}\log L(\theta, \mathbf{x})^2) \tag{3.10}$$

Gdzie $L(\theta, \mathbf{x})$ jest funkcją wiarogodności pod warunkiem zrealizowania się próby $\mathbf{x}$. Jako, że próba jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie, informacja Fishera dla próby wyznacza się jako $nI(\theta)$. Natomiast sama macierz informacyjna ma następującą postać:

$$(J_x)_{ij} = E\left(\frac{\partial \log f(x)}{\partial x_i} \cdot \frac{\partial \log f(x)}{\partial x_j}\right) \tag{3.11}$$

Związek, jaki występuje między informacją Fishera, a lokalną jednoznacznością estymatorów opisany jest m.in w książce [13]. Możemy się w niej zapoznać z dowodem twierdzenia, że koniecznym warunkiem istnienia lokalnej jednoznaczności parametru jest możliwość wyznaczenia jego **nieobciążonego i zgodnego** estymatora. Pierwsza własność takiego estymatora gwarantuje nam, że z dokładnością do losowego odchylenia poprawnie zidentyfikowaliśmy parametr lub równoważnie „średnio rzecz biorąc" nie popełniliśmy błędu przy estymacji [1]. Natomiast druga własność zapewnia nam, że przy odpowiednio dużej próbie bezwzględna różnica między wartością estymatora a prawdziwą wartością parametru będzie „nieznaczna". [2]. Zgodność estymatora jest związana z jego efektywnością. Dowód wspomnianego twierdzenia opiera się o tzw. **nierówność Cramera-Rao**, która przedstawia następującą zależność:

$$D^2(ENMW(\theta)) \geqslant (I^(\theta))^{-1} \tag{3.12}$$

Skrót ENMW oznacza estymator nieobciążony o minimalnej wariancji. W ten sposób informacja Fishera zawiera informację na temat dolnej granicy dokładności oszacowania. Porównanie z nią wartości wariancji uzyskanych nieobciążonych estymatorów parametrów niesie nam informację na ile jesteśmy bliscy możliwie dokładnego oszacowania. Oczywiście, im mniejsza wariancja estymatora, tym mniejsze otoczenie prawdziwej wartości parametru, co z kolei implikuje większą szansę na istnienie lokalnej identyfikowalności modelu.

3.3. Estymacja modelu

Z podstawowego kursu statystyki znamy podstawowe metody estymacji parametrów. Są nimi: metoda momentów, metoda najmniejszych kwadratów oraz metoda największej wiarogodności. Każda z nich ma swoje zalety i ułomności w zależności od rozpatrywanego problemu. Ich dokładny opis znacznie przekracza ramy tej pracy, dlatego warto sięgnąć np. do [55] lub [**?**].

Nas najbardziej interesować będzie metoda największej wiarogodności (ang. maximum likelihood). Jak podaje Silvey [55] jest ona stosowana wtedy, gdy znamy postać rozkładu na przestrzeni próbek, ale nadal pozostaje skończona liczba parametrów do oszacowania.

W interesujących nas zagadnieniach proces estymacji jest o wiele bardziej skomplikowany. Przez wiele lat, gdy statystyka nie była wspomagana komputerowo stosowane były inne

[1] Formalnie oznacza to, że wartość oczekiwana obciążenia $E(b) = E(\theta - \hat{\theta}) = 0$

[2] W ścisłym sensie oznacza to, że ciąg estymatorów jest zbieżny według prawdopodobieństwa do prawdziwej wartości parametru

metody estymacji. Jak podaje Wolfe [62, s.330] zauważa, Karl Pearson w 1894 roku zastosował metodę momentów do estymacji parametrów mieszaniny dwóch jednowymiarowych rozkładów normalnych. Przez długi okres zagadnienie wydzielenia składników było jednak problemem czysto matematycznym. Tak było do momentu, dopóki C.R. Rao w 1952 roku nie zaproponował metody największej wiarogodności dla tego samego problemu. Rozwój mocy obliczeniowej komputerów zaczął sprzyjać rozwojowi tej metody estymacji i obecnie zaimplementowana jest w prawie każdym pakiecie statystycznym (m.in w testowanych w ostatnim rozdziale tej pracy pakietach Latent Gold czy środowisku R).

3.3.1. Modele dyskretne

Zachodzą następujące równości, które będą użyteczne przy wyznaczaniu estymatorów. Prawdopodobieństwo brzegowe uzyskania wartości t zmiennej X można wyrazić jako:

$$\pi_t^X = \sum_{i,j,k} \pi_{ijkt}^{ABCX} \tag{3.13}$$

Natomiast dla dowolnej zmiennej obserwowalnej, łączny rozkład wskaźnika i cechy ukrytej wyraża się następująco:

$$\pi_t^X \pi_{it}^{\bar{A}X} = \sum_{i,j,k} \pi_{ijkt}^{ABCX} \tag{3.14}$$

Symbol $\sum_{i,j,k}$ oznacza, że sumowanie odbywa się kolejno po wszystkich indeksach. Z powyższych równości bezpośrednio uzyskujemy postać odpowiadających im estymatorów największej wiarogodności:

$$\begin{aligned}
\hat{\pi}_{ijkt} &= \sum_{t=1}^{T} \hat{\pi}_{ijkt}^{ABCX} = \sum_{t=1}^{T} \hat{\pi}_t^X \hat{\pi}_{it}^{\bar{A}X} \hat{\pi}_{jt}^{\bar{B}X} \hat{\pi}_{kt}^{\bar{C}X} \\
\hat{\pi}_{ijkt}^{ABC\bar{X}} &= \frac{\hat{\pi}_{ijkt}^{ABCX}}{\hat{\pi}_{ijkt}} \\
\hat{\pi}_t^X &= \sum_{i,j,k} p_{ijkt} \hat{\pi}_{ijkt}^{ABCX} \\
\hat{\pi}_{it}^{\bar{A}X} &= \frac{\sum_{j,k} p_{ijkt} \cdot \hat{\pi}_{ijkt}^{ABCX}}{\hat{\pi}_t^X} \\
\hat{\pi}_{jt}^{\bar{B}X} &= \frac{\sum_{i,k} p_{ijkt} \cdot \hat{\pi}_{ijkt}^{ABCX}}{\hat{\pi}_t^X} \\
\hat{\pi}_{kt}^{\bar{C}X} &= \frac{\sum_{i,j} p_{ijkt} \cdot \hat{\pi}_{ijkt}^{ABCX}}{\hat{\pi}_t^X}
\end{aligned}$$

3.3.2. Modele ciągłe

Załóżmy, że z populacji pobrano próbę wielkości n. Zakładamy, że próba jest ciągiem niezależnych m-wymiarowych wektorów losowych o jednakowym rozkładzie. Obserwacja o numerze k reprezentowana jest przez realizację wektora $x_k = (X_{1k}, X_{2k}, ..., X_{mk})$.

Prawdopodobieństwo, że dana obserwacja została wygenerowana przez rozkład j lub równoważnie, że dany profil należy do skupienia o numerze j wyraża się wzorem:

$$P(j|\mathbf{x}) = \frac{P(\mathbf{x}|j)P(j)}{P(\mathbf{x})} = \frac{\lambda_j f(\mathbf{x}|\theta_j)}{f(\mathbf{x})} \tag{3.15}$$

Logarytm funkcji wiarogodności dla próby o liczebności n wyraża się następującym wzorem:

$$l(\theta, \mathbf{x}) = \sum_{i=1}^{n} \log \sum_{j=1}^{k} \lambda_j f(\mathbf{x}_i|\theta_j) \tag{3.16}$$

Do oszacowania jest zatem zbiór parametrów określających udział poszczególnych składników: $(\lambda_1, \lambda_2, ..., \lambda_j)$ oraz definiujących warunkowy łączny rozkład wskaźników: $\theta_1, \theta_2, ..., \theta_j$. Maksymalizacja funkcji odbywać się będzie przy pewnych ograniczeniach. W naszym przypadku tym ograniczniem jest fakt, że $\sum i = 1^j \lambda_i = 1$. Operacja ta nazywa się poszukiwaniem maksimów warunkowych na zbiorze ograniczonym przy użyciu mnożników Langrange'a. Tak zmodyfikowany logarytm funkcji wiarogodności wyraża się następująco:

$$l(\theta, \mathbf{x}, \omega) = \sum_{i=1}^{n} \log \sum_{j=1}^{k} \lambda_j f(\mathbf{x}_i|\theta_j) - \omega(\sum_{j=1}^{k} \lambda_j - 1) \tag{3.17}$$

Warunkiem koniecznym, aby punkt był ekstremum danej funkcji jest zerowanie się jej pochodnej w tym punkcie. Zatem szukanie maksimów funkcji wiarogodności odbywa się za pomocą przyrównania jej pochodnych cząstkowych do zera:

$$\begin{aligned}
\frac{\partial l(\theta, \mathbf{x}, \omega)}{\partial \lambda_1} &= \sum_{i=1}^{n} \frac{f(\mathbf{x}_i|\theta_1)}{f(\mathbf{x}_i)} - \omega = 0 \\
\frac{\partial l(\theta, \mathbf{x}, \omega)}{\partial \lambda_2} &= \sum_{i=1}^{n} \frac{f(\mathbf{x}_i|\theta_2)}{f(\mathbf{x}_i)} - \omega = 0 \\
&\dots \\
\frac{\partial l(\theta, \mathbf{x}, \omega)}{\partial \lambda_k} &= \sum_{i=1}^{n} \frac{f(\mathbf{x}_i|\theta_k)}{f(\mathbf{x}_i)} - \omega = 0
\end{aligned}$$

Co wynika z faktu, że $\mathbf{x}_i = \sum_{j=1}^{k} \lambda_j f(\mathbf{x}_i|\theta_j)$. Pochodne cząstkowe względem drugiej grupy parametrów opisane są następującymi równaniami:

$$\begin{aligned}
\frac{\partial l(\theta, \mathbf{x}, \omega)}{\partial \theta_1} &= \sum_{i=1}^{n} \frac{\lambda_1}{f(\mathbf{x})} \frac{\partial f(\mathbf{x}_i|\theta_1)}{\partial \theta_1} = 0 \\
\frac{\partial l(\theta, \mathbf{x}, \omega)}{\partial \theta_2} &= \sum_{i=1}^{n} \frac{\lambda_2}{f(\mathbf{x})} \frac{\partial f(\mathbf{x}_i|\theta_2)}{\partial \theta_2} = 0 \\
&\dots \\
\frac{\partial l(\theta, \mathbf{x}, \omega)}{\partial \theta_k} &= \sum_{i=1}^{n} \frac{\lambda_k}{f(\mathbf{x})} \frac{\partial f(\mathbf{x}_i|\theta_k)}{\partial \theta_k} = 0
\end{aligned}$$

Wolfe stosując tę metodę, uzyskał postać estymatorów największej wiarogodności dla szukanych parametrów proporcji:

$$\hat{\lambda}_j = \frac{1}{n}\sum_{i=1}^{n}\hat{P}(j|\mathbf{x}) \tag{3.18}$$

Jeśli chodzi o estymatory parametrów rozkładów warunkowych θ to ich dokładna postać zależy od specyfiki założonego modelu. Dla przykładu, jeśli założymy wielowymiarową normalność rozkładów wówczas $\theta = (\mu_1, \mu_2, ..., \mu_j, \Sigma_1, \Sigma_2, ..., \Sigma_j)$ gdzie współrzędne na miejscach od 1 do j definiują p - wymiarowe wektory średnich, zaś współrzędne od $j+1$ do $2j$ oznaczają $p \times p$ wymiarowe macierze kowariancji w każdej klasie.

Dla przypadku ogólnego podstawiając w powyższych równaniach $\frac{P(j|x)}{f(x|\theta_j)}$ w miejsce $\frac{\lambda_j}{f(x)}$ (równość wynika bezpośrednio z równania Bayesa), uzyskamy równanie postaci:

$$\frac{\partial l(\theta, \mathbf{x}, \omega)}{\partial \theta_j} = \sum_{i=1}^{n} P(j|x)\frac{\partial \log f(x)}{\partial \theta_j}$$

Jak łatwo zauważyć, równania dla estymatorów θ są po prostu średnimi ważonymi (z wagami równymi warunkowym prawdopodobieństwom przynależnośći do danej klasu) pojedynczych równań w przypadku, gdyby łączny rozkład wskaźników generowany był tylko przez jedną klasę ukrytą.

3.3.3. Mieszanina rozkładów normalnych

Agitację za stosowaniem metody największej wiarogodności przy założeniu mieszaniny gęstości rozkładów można znaleźć między innymi w [12]. Również([48], s.319) podaje, że analiza skupień jest właśnie jedną z praktycznych sytuacji, gdzie liczba składników mieszaniny rozkładów odpowiada liczbie rozłącznych skupień, które należy wyznaczyć.

Wolfe: Mimo, że inne procedury mogą być efektywne w szczególnych przypadkach, metodę największej wiarogodności można łatwo uogólnić. Była ona pomijana ze względu na olbrzymi rozmiar obliczeń wymaganych do rozwiązania układu równań. Prześledźmy, w jaki sposób może być stosowana do zagadnienia analizy skupień.

Niech $f_1(\mathbf{x}, \theta_1), f_2(\mathbf{x}, \theta_2), \ldots, f_j(\mathbf{x}, \theta_j)$ będą rozkładami prawdopodobieństwa dla m - wymiarowych wektorów losowych $\mathbf{x} = (X_1, X_2, ..., X_m)$. Ponadto zakładamy, że rozkłady prawdopodobieństwa wyrażają "porządnymi" funkcjami, w tym sensie, że są co najmniej dwukrotnie różniczkowalne względem wektora parametrów θ.

Odpowiednie estymatory kolejno, dla parametrów proporcji składników mieszaniny, wartości oczekiwanej i macierzy kowariancji wyrażają się odpowiednio:

$$\hat{\lambda}_j = \frac{1}{n}\sum_{i=1}^{n}\hat{P}(j|\mathbf{x}_i) \tag{3.19}$$

$$\hat{\mu}_j = \frac{1}{n\hat{\lambda}_j}\sum_{i=1}^{n}\hat{P}(j|\mathbf{x}_i)\cdot x_{ij} \tag{3.20}$$

$$\hat{\Sigma}_j = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mu_j)(\mathbf{x}_i - \mu_j)^T\hat{P}(j|\mathbf{x}_i)\cdot x_{ij} \tag{3.21}$$

Alternatywnym sposobem estymacji modelu z funkcjami gęstości jest metoda zaproponowana przez (Banfield, Raftery) i opisana później w [25] o nazwie klasyfikacja największej wiarogodności (ang. classification maximum likelihood). W przeciwieństwie do modelu mieszanego, nie zakłada sie, że zaobserwowany profil jest efektem działania różnych gęstości, lecz pochodzi z jednego tylko rozkładu zdeterminowanego przez wartość dodatkowego parametru $\gamma_i \in \{1, 2, ..., G\}$, który jest zdefiniowany następująco: $\gamma_i = j$ jeśli x_i pochodzi z j-tego rozkładu. Funkcja wiarogodności dla n-elementowej próby przy użyciu tej metody daje się zapisać jako:

$$L_c(\Theta, \gamma_1, \gamma_2, \cdots, \gamma_k|\mathbf{x}) = \prod_{i=1}^{n} f_{\gamma_i}(x_i|\theta_{\gamma_i}) \tag{3.22}$$

3.3.4. Algorytm EM

Znalezienie rozwiązań dla powyższych układów równań, mimo ich eleganckiego zapisu, zwykle nie jest łatwe. Można wymienić co najmniej trzy zasadnicze problemy.

Po pierwsze, prawa strona układów równań zawiera element $\hat{P}(j|s)$, który jest jedynie estymatorem, a nie prawdziwą wartością prawdopodobieństwa przynależności obserwacji do danej klasy. Gdybyśmy znali jego wartość, cała procedura analizy skupień była bezcelowa.

Po drugie, charakter związku między wskaźnikami, a cechą ukrytą jest nieliniowy, co powoduje, że układ równań może mieć wiele rozwiązań.

Po trzecie, w wielu przypadkach przedstawione układy równań muszą być rozwiązane numerycznie. Bezpośrednia maksymalizacja funkcji wiarogodności jest problemem o tyle złożonym ponieważ prawa strona równania zawiera znak sumy pod logarytmem [32].

Algorytm, który zostanie za chwilę przedstawiony rozwiązuje zadanie estymacji w sposób, który nie tylko bazuje na inteligentnym przeformułowaniu problemu, ale też to przeformułowanie ma mocną podbudowę teoretyczną, o której mówiliśmy w pierwszej części tego rozdziału.

Widzimy, że obie strony równania zależą od estymatorów parametrów uzyskanych w poprzednich krokach. Sugeruje to zastosowanie pewnej procedury iteracyjnej. Metoda, którą proponuje Goodman jest uniwersalna dla szerokiej klasy modeli z cechą ukrytą. Jego pomysł, który został szczegółowy opisany dla zmiennych dyskretnych pojawi się w przypadku modelu ciągłego jako algorytm EM.

Opis i rozwój algorytmu

Nazwa algorytmu pochodzi od pierwszych liter dwóch na przemian wykonywanych kroków. Pierwszy z nich to **Expectation** i polega na liczeniu wartości oczekiwanej pewnej statystyki, a drugi to **Maximalization**, podczas którego maksymalizacji poddawana jest funkcja wiarogodności.

Oficjalnie algorytm EM liczy sobie ponad 30 lat. Wiele źródeł traktuje pracę Dempstera, Lairda i Rubina [21] z 1977 roku jako początek prac nad algorytmem. Niewątpliwie, ten powszechnie cytowany artykuł, oprócz formalnych pojęć zawiera dokładne wprowadzenie ilustrowane elementarnymi przykładami. Sami autorzy powołują się na wcześniejszych autorów, a Xiao, Dyk [63] w swoim tekście napisanym 20 lat po ukazaniu się pracy [21] porównują EM do pieśni ludowej, która była znana dużo wcześniej zanim po raz pierwszy zarejestrowano jej formalny zapis. Oni również odsyłają czytelnika do propozycji Hartleya z 1958 roku, ale ogólną konkluzja jest parafrazą myśli filozoficznej, że *w końcu wszystkie współczesne prace statystyczne są komentarzami do dzieł sir R.A. Fishera.* Interesującą ilustracją w artykule

jest wykres przedstawiający liczbę prac na temat algorytmu EM w zależności od czasu oraz liczbę cytowań pracy [21], co faktycznie czyni ją klasyczną w swojej kategorii.

Zakres zastosowań algorytmu jest równie szeroki jak dla metody największej wiarogodności. [21] podają kilka możliwości: analiza czynnikowa, analiza wariancji, estymacja uwzględniająca braki danych, rozwiązywanie problemu mieszaniny rozkładów i klasyfikacji. Dla nas najbardziej intresujące będą dwa ostatnie rodzaje aplikacji.

Idea, która stoi za przeformułowaniem problemu jest zgodna z teorią klasy ukrytej. Załóżmy, że każda obserwacja posiada dodatkową zmienną ukrytą γ, której początkowe wartości są nieznane (lub brakujące). Taki proces niekiedy nazywany jest powiększeniem danych (ang. data augmentation - [32]).[3] Zmienną ukrytą jest oczywiście klasyfikacja, wobec czego przyjmuje ona wartości między 0 i 1. Taki zabieg jest równoważny **uzmiennieniu** wektora parametrów λ odpowiedzialnych za proporcję składników mieszaniny. Nowa zmienna ukryta γ_i przyjmuje wartość 1 gdy obserwacja o numerze i należy do skupienia o numerze k. W związku z tym, że obserwacje mogą przynależeć do jednego i tylko jednego skupienia, a ponadto przynależą niezależnie od siebie , γ_i ma wielowymiarowy rozkład Bernoulliego, czyli po prostu rozkład wielomianowy. Odpowiada to sytuacji umieszczenia jednej kuli w G ponumerowanych komórkach, gdzie prawdopodobieństwo trafienia do k jest proporcjonalne do jej wielkości. Czyli formalnie:

$$\gamma_i \sim Mult(1, \lambda_1, \lambda_2, ..., \lambda_G)$$

Mając określony rozkład nowej zmiennej, możemy zapisać warunkowy rozkład wskaźników względem zmiennej ukrytej:

$$f(\mathbf{x}, \gamma_i) = \prod_{k=1}^{G} f(\mathbf{x}|\theta_k)^{\gamma_{ik}} \tag{3.23}$$

Przyjrzyjmy się teraz, w jaki sposób można zmodyfikować funkcję wiarogodności dla n - elementowej próby. Jej oryginalna postać jest następująca:

$$L(\theta, \lambda|\mathbf{x}) = \prod_{i}^{n} \sum_{j=1}^{k} \lambda_j f_j(x_i|\theta_j) \tag{3.24}$$

Po zlogarytmowaniu stronami uzyskujemy:

$$l(\theta, \lambda|\mathbf{x}) = \sum_{i}^{n} \log \sum_{j=1}^{k} \lambda_k f_j(x_i|\theta_j)$$

Wykonując podstawienie z 3.23 otrzymujemy nową postać logarytmu funkcji wiarogodności, którego maksymalizacja jest o wiele mniej złożona numerycznie niż wyjściowa:

$$l(\theta, \lambda, \gamma_{ij}|\mathbf{x}) = \sum_{i}^{n} \sum_{j=1}^{k} \gamma_{ik}(\log \lambda_j f(x_i|\theta_j))$$

Mając tak zapisany logarytm funkcji wiarogodności możemy przystąpić do analizy algorytmu EM.

[3] Alternatywny sposób modyfikacji funkcji wiarogodności można znaleźć w [42, s.359] Autorzy pokazują, że problem maksymalizacji można przeprowadzić niezależnie dla każdej grupy parametrów

Przebieg

Podstawowe intuicje związane z przebiegiem algorytmu oraz graficzną ilustrację kolejnych jego kroków w odniesieniu do mieszanin rozkładów normalnych można znaleźć w Dellaert [20]. Warto zapoznać się również z dydaktycznym tekstem I.Dinova, jednego z autorów witryny Statistic Online Computational Resource (http://www.socr.ucla.edu/), która stanowi interaktywną bazę narzędzi i wizualizacji danych dla różnych modeli. W samym tekście, oprócz dokładnego opisu metody wiarogodności i samego algorytmu EM, można zobaczyć proces estymacji dla rodziny rozkładów normalnych i Poissona oraz prześledzić kolejne kroki algorytmu dla przypadków 1, 2 i 3 - wymiarowych. Czytelnik zainteresowany wyższym pułapem ścisłości powinien zapoznać się z pracą Rednera i Walkera [53].

Algorytm **EM** działa w czterech następujących krokach:

1. Zainicjuj parametry początkowe λ oraz θ.
2. **Krok E**: Dla każdej obserwacji oblicz wartość oczekiwaną γ przynależności do danej klasy, pod warunkiem ustalonych pozostołych parametrów
3. **Krok M**: Dla danej wartości oczekiwanej γ zmaksymalizuj funkcję wiarogodności i zwróć estymatory $\hat{\lambda}$ oraz $\hat{\theta}$, które ją maksymalizują
4. Estymatory uzyskane w kroku 3 podstaw do kroku 2. Kroki 2 i 3 wykonuj naprzemiennie do uzyskania zbieżności

Słowo komentarza. Pierwszy krok polega na „odgadnięciu" lub wylosowaniu wyjściowych parametrów. Zauważmy, że podobny zabieg stosowany jest również w metodzie K-średnich. [4]. Analogicznie niestety, nie jest znana uniwersalna poprawna metoda wyznaczania punktów startowych.

Drugi krok polega policzeniu warunkowego prawdopodobieństwa przynależnośći do klasy, czyli podobnie jak w ogólym przypadku, korzystamy z reguły Bayesa:

$$P(j|x) = E(\gamma_{ik}) = \frac{\lambda_k}{f}(\mathbf{x}_i|\theta_k)\sum_{j=1}^{G}\lambda_j f(\mathbf{x}_i|\theta_j) \tag{3.25}$$

Trzeci krok polega na optymalizacji zagadnienia największej wiarogodności i aktualizacji wyjściowych parametrów. Przeliczone parametry służą do obliczenia wartości oczekiwanej γ, dla wszystkich obserwacji, co jest równoważne przemieszczaniu ich między skupieniami.

Po skończonym procesie iteracji uzyskujemy optymalne rozwiązanie będące estymatorem największej wiarogodności dla wektora parametrów. Trzeba jednak zadać sobie dwa pytania:

1. Co należy rozumieć poprzez optymalność rozwiąznia?
2. W jakim sensie jest ono optymalne?

Zacznijmy od pierwszej kwestii. Jeśli przez γ^*_{ik} oznaczymy wartość oczekiwaną γ_{ik}, która maksymalizuje funkcję wiarogodności, to klasyfikacją obserwacji o numerze i nazwiemy indeks j, dla którego γ^*_{ik} jest największa, przy czym maksymalizacja odbywa się po wszystkich indeksach skupień. Wówczas niepewność związaną z przypisaniem obserwacji do właściwego skupienia można wyrazić poprzez $1 - \gamma^*_{ik}$. Zatem można powiedzieć, że dobra klasyfikacja to

[4] Co więcej [32] pokazują, że K-średnich jest tylko szczególnym przypadkiem algorytmu EM

taka, która maksymalizuje prawdopodobieństwo przynależności do jednej klasy i minimalizuje dla wszystkich pozostałych. Jak łatwo zauważyć, jest to równoważne minimalizacji warunkowej entropii zmiennej γ dla każdej obserwacji. Jak zauważają Arminger i Kusters [5, s.377] jest to zgodne z ogólnym postulatem konstrukcji modeli z cechą ukrytą, aby każdy wskaźnik (tutaj każda obserwacja) związany był tylko z jedną cechą ukrytą (tutaj z jednym skupieniem). Jako analogię podają postulaty "prostej struktury" Thurstone'a z 1947 roku dla modelu czynnikowego. Należy jednak pamiętać, że w analizie czynnikowej prosta struktura była efektem pewnych zabiegów algebraicznych (rotacje macierzy struktury) mających na celu wizualizację pewnego optymalnego rozwiązania. Natomiast w zagadnieniu analizy skupień, osiągnięcie prostej struktury jest postulatem dla optymalnego rozwiązania.

Problem stabilności lub globalnej optymalności jest nieco bardziej skomplikowany. Pamiętamy, że podobne algorytmy (w tym K-średnich) wykazują wrażliwość na zmianę warunków początkowych. Podobnie jest w przypadku algorytmu EM - mamy gwarancję, że uzyskane rozwiązanie jest optymalne jedynie w lokalnym sensie. Na szczęście EM jest dość stabilnym algorytmem - niewielkie zaburzenia warunków początkowych prowadzą do nieistotnych różnic w zwracanym rozwiązaniu.

Jednym z bardziej powszechnych zastosowań algorytmu EM w kontekście analizy skupień jest identyfikacja parametrów mieszaniny rozkładów. Zwykle przyjmuje się, że mamy do czynienia z rozkładami normalnymi lub w ogólnym przypadku należą one do rodziny rozkładów wykładniczych. Ogólną postać funkcji gęstości dla tej klasy modeli można znaleźć np. w [**?**]:

$$f_\theta(x) = \exp\{\sum_{c_j=1}^{k} c_j(\theta)S_j(x) - b(\theta)\} \cdot h(x) \tag{3.26}$$

W powyższym wzorze $S_1(x), S_2(x), \cdots, S_k(x)$ są liniowo niezależne, natomiast $\{c_1(\theta), c_2(\theta), \cdots, c_k(\theta) : \theta \in \Theta\} \in R^k$

Po zlogarytmowaniu stronami powyższego równania otrzymujemy:

$$\log f_\theta(x) = \mathbf{S(x)}^T \cdot \mathbf{c}(\theta) - b(\theta) + h'(x) \tag{3.27}$$

Na podstawie twierdzenia dla rodziny wykładniczej ([**?**, s.29]), $\mathbf{S(x)}$ jest k-wymiarową statystyką dostateczną. Stosując tę terminologię Navidi ([51, s.29]) pokazał, że algorytm EM można wyrazić rekurencyjnie za pomocą wyżej zdefiniowonych pojęć. Załóżmy, że wykonaliśmy pewną liczbę iteracji równą p otrzymując wektor estymatorów θ_p:

1. Krok **E**: Oblicz warunkową wartość oczekiwaną statystyki dostatecznej: $E(\mathbf{S}|y, \theta_p)$
2. Krok **M**: W 3.27 dokonaj podstawienia $\mathbf{S} := E(\mathbf{S}|y, \theta_p)$, a następnie znajdź wartość θ_{p+1}, które maksymalizuje prawą stronę tego równania.

W tym samym tekście możemy również poznać dwie przydatne własności algorytmu EM:

1. Jeśli θ jest parametrem (wektorem parametrów) rzeczywistym, wówczas ciąg oszacowań parametrów zbiega monotonicznie do estymatorów największej wiarogodności.
2. W każdej iteracji rośnie wartość funkcji wiarogodności.

Dowody obu własności można znaleźć właśnie w [51], choć szczegółowe wyjaśnienie znajduje się również w [32]. W następnej sekcji postaramy się zademonstrować działanie algorytmu metodą "krok po kroku". Na początku przedstawimy procedurę autorstwa Goodmana dla modeli klasy ukrytej, a później zilustrujemy przebieg algorytmu dla mieszanek rozkładów normalnych.

3.3.5. Przykład

Rozpatrzmy klasę modeli zaproponowaną przez Goodmana tj. z czterema zmiennymi obserwowalnymi A, B, C, D. Dla prosty, przyjmijmy, że mają one charakter dychotomiczny. Ich łączny rozkład częstości w próbie liczącej 1000 obserwacji przedstawia się następująco:

Tabela 3.2: Rzeczywiste rozkłady liczebności profili

profil ABCD	liczebność
„1111”	150
„1110”	100
„1101”	100
„1011”	50
„0111”	50
„1100”	50
„1010”	50
„0110”	50
„1001”	50
„0101”	50
„0011”	50
„1000”	50
„0100”	50
„0010”	50
„0001”	50
„0000”	50

Zakładamy, ze w populacji istnieje jedna zmienna ukryta, również o charakterze dychotomicznym, która generuje rozkład wskaźników z nieznanymi prawdopodobieństwami warunkowymi. Razem z rozkładem częstości tej zmiennej, stanowią one nieznany wektor parametrów, który należy oszacować. Upewnijmy się, że model jest identyfikowalny sprawdzając warunki podane przez Gibsona oraz Goodmana. W naszym przypadku $I \cdot J \cdot K \cdot = 16$, a $(I + J + K + L - 3) \cdot T = 10$ zatem mamy szansę na lokalną identyfikowalność parametrów.

Łączny rozkład wszystkich zmiennych, bez wartości parametrów przedstawia się następująco:

Pierwszy krok algorytmu polega na zainicjowaniu wartości początkowych nieznanych parametrów. Przyjmijmy ich następujące wartości:

Kolejny krok algorytmu zilustrujemy za pomocą skróconej wersji tabeli:

Naturalnie, w tym konkretnym przypadku $\pi_2^X = 1 - \pi_1^X$ oraz np. $\pi_{01}^{\bar{A}X} = 1 - \pi_{11}^{\bar{A}X}$. Dwie pierwsze kolumny zawierają estymatory łącznego rozkładu częstości wskaźników w każdej z klas. Następna kolumna niesie informację na temat empirycznego rozkładu wskaźników. Kolejna powstała przez dodanie wartości dwóch pierwszych kolumn otrzymując w ten sposób wartość estymatora łącznego rozkładu wskaźników. Dwie następne kolumny reprezentują estymowane prawdopodobieństwa przynależności profilu do danej klasy ukrytej i zostały wyliczone na podstawie reguły Bayesa poprzez podzielenie odpowiednio - pierwszej kolumny przez czwartą i drugiej przez piątą. Dwie ostatnie kolumny stanowią robocze dane niezbędne do obliczenia parametrów w następnym kroku algorytmu. Przedostatnia kolumna przedstawia składniki wzoru na prawdopodobieństwo całkowite, że obserwacja należy do pierwszej klasy ukrytej. Dokładniej, zsumowanie tych elementów prowadzi do wyznaczenia estymatora częstości klasy ukrytej, który będzie wykorzystany do estymacji pozostałych częstości w następnym kroku.

Tabela 3.3: Rzeczywiste częstości profili i model z brakującymi parametrami

	klasa		
zmienna	I	II	
X	.	.	
A	.	.	
B	.	.	
C	.	.	
D	.	.	
profil			π_{ijkl}^{ABCD}
"1111"	.	.	0.15
"1110"	.	.	0.1
"1101"	.	.	0.1
"1011"	.	.	0.05
"0111"	.	.	0.05
"1100"	.	.	0.05
"1010"	.	.	0.05
"0110"	.	.	0.05
"1001"	.	.	0.05
"0101"	.	.	0.05
"0011"	.	.	0.05
"1000"	.	.	0.05
"0100"	.	.	0.05
"0010"	.	.	0.05
"0001"	.	.	0.05
"0000"	.	.	0.05

Tabela 3.4: Zainicjowane parametry modelu

	klasa		
zmienna	I	II	
X	0.4	0.6	
A	0.8	0.1	
B	0.9	0.4	
C	0.8	0.1	
D	0.4	0.3	
x			p(x)
"1111"	.	.	0.15
"1110"	.	.	0.1
"1101"	.	.	0.1
"1011"	.	.	0.05
"0111"	.	.	0.05
"1100"	.	.	0.05
"1010"	.	.	0.05
"0110"	.	.	0.05
"1001"	.	.	0.05
"0101"	.	.	0.05
"0011"	.	.	0.05
"1000"	.	.	0.05
"0100"	.	.	0.05
"0010"	.	.	0.05
"0001"	.	.	0.05
"0000"	.	.	0.05

Tabela 3.5: Add caption

	I	II	π_{ijkl}^{ABCD}	$\hat{\pi}_{ijkl}^{ABCD}$	$\hat{\pi}_{ijkl1}^{ABCD\bar{X}}$	$\hat{\pi}_{ijkl2}^{ABCD\bar{X}}$	$\hat{\pi}_{ijkl1}^{ABCD\bar{X}} \cdot \pi_{ijkl}^{ABCD}$	$\hat{\pi}_{ijkl2}^{ABCD\bar{X}} \cdot \pi_{ijkl}^{ABCD}$
"1111"	$\pi_1^X \pi_{11}^{\bar{A}X} \pi_{11}^{\bar{B}X} \pi_{11}^{\bar{C}X} \pi_{11}^{\bar{D}X}$	$\pi_2^X \pi_{12}^{\bar{A}X} \pi_{12}^{\bar{B}X} \pi_{12}^{\bar{C}X} \pi_{12}^{\bar{D}X}$						
	0.092	0.001	0.15	0.093	0.992	0.008	0.149	0.001
...	...	...	...	...	...	...	...	...
"0000"	$\pi_1^X \pi_{01}^{\bar{A}X} \pi_{01}^{\bar{B}X} \pi_{01}^{\bar{C}X} \pi_{01}^{\bar{D}X}$	$\pi_2^X \pi_{02}^{\bar{A}X} \pi_{02}^{\bar{B}X} \pi_{02}^{\bar{C}X} \pi_{02}^{\bar{D}X}$						
	0.001	0.204	0.05	0.205	0.005	0.995	0.000	0.050

Tabela 3.6: Wartości estymatorów parametrów w kolejnej iteracji

	I	II				
X	0.566	0.434				
A	0.824	0.308				
B	0.781	0.365				
C	0.755	0.283				
D	0.595	0.491	π_{ijkl}^{ABCD}	$\hat{\pi}_{ijkl}^{ABCD}$	$\hat{\pi}_{ijkl1}^{ABCD\bar{X}}$	$\hat{\pi}_{ijkl2}^{ABCD\bar{X}}$
"1111"	0.164	0.007	0.15	0.170	0.960	0.040
"1110"	0.111	0.007	0.1	0.118	0.941	0.059
"1101"	0.053	0.017	0.1	0.070	0.756	0.244
"1011"	0.046	0.012	0.05	0.058	0.796	0.204
"0111"	0.035	0.015	0.05	0.050	0.697	0.303
"1100"	0.036	0.018	0.05	0.054	0.670	0.330
"1010"	0.031	0.012	0.05	0.043	0.719	0.281
"0110"	0.024	0.016	0.05	0.040	0.601	0.399
"1001"	0.015	0.030	0.05	0.045	0.333	0.667
"0101"	0.011	0.039	0.05	0.050	0.227	0.773
"0011"	0.010	0.026	0.05	0.036	0.270	0.730
"1000"	0.010	0.031	0.05	0.041	0.246	0.754
"0100"	0.008	0.040	0.05	0.048	0.162	0.838
"0010"	0.007	0.027	0.05	0.034	0.195	0.805
"0001"	0.003	0.067	0.05	0.070	0.045	0.955
"0000"	0.002	0.070	0.05	0.072	0.030	0.970

W porównaniu z estymowanymi wartościami łącznego rozkładu wskaźników obliczanych na podstawie „odgadniętych" parametrów w pierwszym kroku, już w kolejnej iteracji widać znaczną poprawę tj. zbliżoną wartość estymatorów do prawdziwych wartości parametrów. Tempo zbieżności estymatorów można zilustrować na następującym wykresie:

(tu convergence.eps)

3.3.6. Mocne i słabe strony metod

Niewątpliwie najczęściej wymienianą zaletą algorytmu jest jego prostota. Chodzi tu przede wszystkim o przejrzystość jego działania, co powoduje, że jest on intuicyjnie zrozumiały. Natomiast z punktu widzenia programistycznego jest on również łatwy w implementacji.

Z praktycznego punktu widzenia najbardziej interesującą cechą jest jego zbieżność bez względu na warunki początkowe. HTF podają przystępny dowód oparty na nierówności Jensena wyjaśniający, czemu algorytm EM nigdy nie dopuszcza spadku wartości funkcji wiarogodności w kolejnych krokach. Z kolei z faktu, że jej wartości są niemalejące względem czasu i ograniczone przez globalne maksimum wynika zbieżność do granicznej wartości.

Podstawowym zarzutem stawianym algorytmowi EM jest jego powolność ([25, 26, 63]), co oznacza, że wymaga dużej liczby iteracji, aby osiągnąć wymagany próg zbieżności. Algorytm może działać wolniej w dwóch przypadkach. Po pierwsze, gdy skupienia nie są dość dobrze odseparowane. Po drugie, gdy punkty startowe zostały dobrane dość niefortunnie. Informacja na temat wyjątkowo powolnej zbieżności może być niezwykle użyteczna do oceny segmentowalności zbioru, pod warunkiem, że wyeliminujemy pozostałe niekorzystne czynniki (np. starannie dobierzemy warunki początkowe). Graficzną ilustrację tempa zbieżności do

estymatora NW wraz z komentarzem można znaleźć w pracy Navidiego [51].

W [63] kilka pomysłów na poprawienie szybkości.

Kolejną wadą jest bezradność algorytmu w obliczu zbyt dużej liczby parametrów. Liczba prawdopodobieństw warunkowych, za które odpowiadają parametry θ jest równa iloczynowi liczby wskaźników i wartości cechy ukrytej. Dla dużej liczby skupień algorytm jest mało praktyczny ([26]). Na szczęście, w wielu przypadkach modelowania dąży się do najprostszego rozwiązania tj. takiego, które potrafi opisać zjawisko za pomocą możliwie małej liczby parametrów (ang. parsimony).

Ostatnim poważnym niedomaganiem algorytmu jest brak odporności na złą określoność macierzy kowariancji między wskaźnikami. Taka macierz jest źle określona, gdy jej wyznacznik jest bliski równy zero. Z praktycznego punktu widzenia, biorąc pod uwagę lokalną niezależność wskaźników (co oznacza, że macierze kowariancji są diagonalne), wyznacznik dąży do zera, gdy co najmniej jeden z elementów na przekątnej dąży do zera. To z kolei równoważne jest sytuacji, gdy wariancja pewnego wskaźnika jest bliska zeru.

Rozdział 4

Konfirmacyjny model analizy skupień

> Pomimo, że może istnieć teoria, która rozstrzygnie zagadnienie liczby skupień, to w przypadku eksploracyjnym gdzie liczba składników mieszaniny jest nieznana, nie ma ogólnej zgody co do tego jak tę liczbę estymować.

Możliwość testowania liczby klas niewątpliwie można uznać za przełomowy moment w rozwoju metodologii analizy skupień. Z jednej strony rozpoczyna nowy, konfirmacyjny rozdział w jej historii, z drugiej zaś uzupełnia podjętą wcześniej koncepcję modelowania. Wyznaczanie liczby skupień za pomocą metod statystycznych jest ważnym, ale niezwykle trudnym zagadnieniem. Przykłady sytuacji, gdzie narzucające się standardowe metody nie mają zastosowania, zostaną zaprezentowane w dalszej części pracy. Problem zostanie omówiony z dwóch perspektyw - bayesowskiej i klasycznej. Motywacja dla równoległego omówienia tych dwóch interpretacji stanie się jasna w momencie, gdy będziemy przedstawiać kryteria wyboru najlepszego modelu.

Ogólna idea testowania liczby klas ukrytych może być opisana następująco. Na początku sprowadzamy wybór liczby k skupień do wyboru pewnego konkretnego modelu M_k (por. [19, s.296]). Na razie zakładamy jednoznaczność modelu ze względu na liczbę skupień. Dla każdej takiej liczby istnieje dokładnie jeden model, który jest najlepszy. Kryterium optymalności może być zdefiniowanie arbitralnie, jednak w naszym przypadku będziemy je rozumieć poprzez posiadanie największej wiarogodności.

W ten sposób każdy z modeli $M_1, M_2, ..., M_G$, który opisywany jest przez zestaw parametrów $\Theta_1, \Theta_2, ..., \Theta_G$, zawiera w indeksie zakładaną liczbę skupień. Na pytanie, w jaki sposób można porównywać modele odpowiemy z dwóch perspektyw.

4.1. Podejście bayesowskie

Nazwa metody pochodzi od nazwiska Bayesa, a dokładniej jego wzoru wiążącego prawdopodobieństwa warunkowe. Obszerne wprowadzenie do metod bayesowskich można znaleźć w każdym specjalistycznym podręczniku statystyki (zob. [**?**, 55]). Tutaj zostaną one przedstawione jedynie w zarysie.

Główna różnca między podejściem częstościowym (ang. frequentist) i bayesowskim (ang. bayesian) polega na tym, że to ostatnie rozróżnia dwa rodzaje prawdopodobieństwa: **a priori** (ang. prior likelihood) i **a posteriori** (ang. posterior likelihood). Zresztą, sama koncepcja prawdopodobieństwa (zauważmy, że nie pojawia się tutaj termin *probability* lecz właśnie *li-*

kelihood) jest inaczej formułowana w obu perspektywach. W podejściu częstościowym, prawdopodobieństwo jest rozumiane jako *oczekiwany* wynik, gdy eksperyment powtarzany jest *odpowiednio wiele* razy. W podejściu bayesowskim mamy do czynienia raczej ze *stopniem przekonania* (równoważnie ufnością), że dany wynik jest prawdopodobny. Stąd często mówi się o *prawdopodobieństwie subiektywnym*. Dodatkowo, w tej perspektywie dopuszczalne jest określenie prawdopodobieństwa dla parametru i całego modelu, co dla pierwszej jest bezcelowe. W podejściu częstościowym parametr jest pewną stałą, więc nie można określić dla niego prawdopodobieństwa. Jednak, jak podaje [55] *tak określone prawdopodobieństwo nie jest żadnym nowym obiektem matematycznym*. Analogicznie, zostają zachowane znane nam zależności między odpowiednimi wielkościami. Dokładniej, jeśli $\pi(\theta)$ oznacza rozkład parametru a priori, wówczas rozkład łączny wskaźników x i θ obliczany jest ze wzoru:

$$p(x, \theta) = \pi(\theta) \cdot p(x|\theta) \tag{4.1}$$

Podobnie, rozkład a posteriori parametru, wyznaczamy z reguły Bayesa:

$$p(\theta|x) = \pi(\theta) \cdot \frac{p(x|\theta)}{p(x)} = \pi(\theta) \cdot \frac{p(x|\theta)}{\int_\Theta \pi(\theta) p(x|\theta), d\theta} \tag{4.2}$$

Całka w mianowniku w ostatnim elemencie prawej strony równości oznacza brzegowy łączny rozkład wskaźników. Całka zastępuje tutaj tradycyjny znak sumy ponieważ nie jesteśmy w stanie określić, czy rozkłady $\pi(\theta_i)$ wyrażają się przez gęstość czy funkcję prawdopodobieństwa - stąd ogólna postać całkowa.

Po tym krótkim wstępie możemy wrócić do naszego problemu wyboru liczby klas. Jeszcze raz, załóżmy, że mamy zbiór M_i, $i = 1, 2, ..., G$ kandydatów na modele i odpowiadające im parametry θ_i. Chcemy wybrać najlepszy z nich. Dla każdego z nich mamy zdefiniowane prawdopodobieństwo a priori wyrażone przez $P(M_i)$. Dla każdego z nich możemy policzyć prawdopodobieństwo a posteriori względem uzyskanych danych w próbie:

$$p(M_i|x) = \frac{p(x|M_i) \cdot p(M_i)}{p(x)} \tag{4.3}$$

Do porównania dwóch konkurencyjnych modeli wykorzystuje się iloraz ich prawdopodobieństw a posteriori, który nazywany jest również stosunkiem szans (ang. odds-ratio).

$$\frac{p(M_0|x)}{p(M_1|x)} = \frac{p(M_0)}{p(M_1)} \cdot \frac{p(x|M_0)}{p(x|M_1)} \tag{4.4}$$

Ostatni czynnik, $\frac{p(x|M_0)}{p(x|M_1)}$ nazywany jest czynnikiem Bayesa (ang. Bayes factor). Zwykle, zgodnie z koncepcją całkowitej niewiedzy przyjmuje się jednostajny rozkład a priori parametrów, co prowadzi do zanikania pierwszego czynnika po prawej stronie równania.

Z samej postaci ilorazu nie wypływają jednak żadne praktyczne wskazówki, co do wyboru najlepszego modelu. Niezbędne jest pewne oszacowanie wielkości prawdopodobieństwa a posteriori. Hastie i in. [32, s.206] uznają, że rozsądna jest tzw. aproksymacja Laplace'a:

$$\log p(x|M_i) = \log p(x|\hat{\theta}_i, M_i) - \frac{d_i}{2} \log N \tag{4.5}$$

Gdzie $\hat{\theta}_i$ oznacza wektor estymatorów największej wiarogodności, d_i - liczbą parametrów, a N jest liczebnością próby. Jeśli zdefiniujemy funkcję straty jako $-2\log p(x|\hat{\theta}_i, M_i)$, wówczas uzyskamy jedno z częściej stosowanych kryteriów wyboru modelu. Kryterium to nosi nazwę BIC (czasem też BIC-Schwarza) i jest akronimem od **B**ayesian **I**nformation **C**riterion. Zgodnie z definicją:

$$BIC = -2\log L + d\log N \tag{4.6}$$

Gdzie L jest zmaksymalizowaną funkcją wiarogodności dla modelu, d liczbą parameteów, a N liczebnością próby. BIC należy do kryteriów, które regulują relację między dokładnością modelu (funkcja wiarogodności) i jego stopniem skomplikowania (liczba parametrów). Nietrudno zauważyć, że jeśli dane dwa modele tak samo dokładnie odtwarzają łączny rozkład wskaźników, to preferowany będzie model prostszy, o mniejszej wartości BIC.

Wartości BIC różnią się w zależności od charakteru modelu, dlatego do porównywania jego wartości posłużymy się tym, od czego zaczęliśmy - prawdopodobieństwem a posteriori. Po elementarnych przekształceniach otrzymujemy:

$$p(x|M_i) = \frac{e^{-\frac{1}{2}BIC_i}}{\sum_{j=1}^{G} e^{-\frac{1}{2}BIC_j}} \tag{4.7}$$

Alternatywne przybliżenie czynnika Bayesa można znaleźć w pracy Smitha i Spiegelhaltera [**?**]:

$$B_k(x) = \Lambda_k - \left(\frac{3}{2} + \log(pn_{k,k+1})\delta_k\right) \tag{4.8}$$

Gdzie Λ_k jest ilorazem wiarogodności dla testowania modelu M_r przeciw M_{k+1}, δ_k jest liczbą stopni swobody statystyki $-2\log\Lambda$ równą różnicy między liczbą parametrów w obu modelach, p - wymiarem przestrzeni wskaźników, a $n_{k,k+1}$ jest rozmiarem skupienia, które powstaje w wyniku przejścia z modelu o $k+1$ skupieniach do modelu o k skupieniach. Zatem kryterium jest bezużyteczne w momencie, gdy posługujemy się algorytmem nie-hierarchicznym.

Kryterium to posiada jednak dwie podstawowe wady. Po pierwsze, zostało stworzone wyłącznie na potrzeby algomeracyjnych metod hierarchicznych, co niewątpliwie stanowi użyteczną ocenę jakości uzyskanego dendrogramu. Niemniej jednak ogranicza się do testowania wyłącznie par modeli różniących się o jedno skupienie. Co więcej, zakładane jest tutaj zawierane się kolejnych podziałów w zależności od liczby skupień. Innymi słowy, zakłada się, że model o dwóch skupieniach zawiera modele o trzech, czterech itd. skupieniach, co jest spełnione tylko w szczególnych przypadkach. Po trzecie, jak podają autorzy na podstawie badań symulacyjnych, przybliżenie to nie jest dokładne, gdy liczba skupień przekracza 5.

4.2. Podejście klasyczne

Naturalnym podejściem jest wykorzystanie testu opartego na ilorazie wiarogodności - LRT (ang. Likelihood Ratio Test). Szczegółowy opis testu wraz z twierdzeniami granicznymi można znaleźć w [1]. O zastosowaniach w statystyce można przeczytać dodatkowo w [55]. Podobnie jak w przypadku metod bayesowskich, tutaj przedstawimy tylko elementarny wykład. Zaprezentowany poniżej tok rozumowania można znaleźć m.in. w [**?**, rozdział 3]

Przez Θ będziemy rozumieć zbiór wszystkich parametrów opisujących daną rodzinę rozkładów. Będziemy mówić, że hipoteza zerowa H_0 (której odpowiada model M_0) wyróżnia w tej przestrzeni pewien podzbiór. Dopełnienie tego pozbioru wyznacza parametry zgodne z hipotezą konkurencyjną H_1 (i odpowiednim modelem M_1).

Dla obu hipotez możemy wyznaczyć ich wiarogodności (podobnie jak czyniliśmy to z prawdopodobieństwem a posteriori): L_0 i L_1. Są to zmaksymalizowane (względem θ) funkcje wiarogodności odpowiadające poszczególnym modelom. Porównanie tych dwóch wielkości prowadzi do testu ilorazu wiarogodności:

$$\Lambda = \frac{L_1}{L_0} \tag{4.9}$$

Zasada największej wiarogodności mówi, że jeśli $\Lambda > \Lambda^* \geqslant 1$ wówczas należy odrzucić model zgodny z hipotezą zerową. Stała Λ^* wyznaczana jest na podstawie ustalanego poziomu istotności α, tzn. tak, aby warunkowe prawdopodobieństwo, że iloraz wiarogodności będzie większy od Λ_0, gdy prawdziwa jest hipoteza zerowa, było niewiększe od przyjętego poziomu istotności:

$$P_{H_0}(\frac{L_1}{L_0} > \Lambda_0) \leqslant \alpha \tag{4.10}$$

Iloraz wiarogodności jest statystyką, a więc posiada pewien rozkład na przestrzeni prób. Mimo, że w ogólnym przypadku nie jest on znany, to w połowie lat trzydziestych ubiegłego stulecia (dokładnie w roku 1935) Wilks pokazał, że jego pewna modyfikacja, zwana statystyką G^2 zbiega według rozkładu do innej dobrze znanej statystyki. Konkretnie:

$$G^2 = -2\log\Lambda \rightarrow \chi^2_k \tag{4.11}$$

gdzie k jest różnicą między liczbą parametrów w obu modelach. Porównywanie dwóch modeli różniących się liczbą ustalonych (lub wyzerowanych) parametrów sprowadza się zatem do konstrukcji odpowiedniego obszaru krytycznego dla statystyki χ^2, obliczenia stopni swobody k i na podstawie empirycznego ilorazu wiarogodności podjęcia odpowiedniej decyzji.

4.3. Wybór najlepszego modelu

Posługując się kryterium BIC możemy obliczyć prawdopodobieństwa a posteriori dla wszystkich modeli i wybrać ten o największej wartości (lub równoważnie, ten o najmniejszej wartości BIC). Jednak, w sytuacji, gdy porównujemy kilka albo kilkanaście modeli jednocześnie, potrzebujemy uzasadnienia, czy uzyskane różnice są statystycznie istotne. W tym miejscu pojawia się pierwsza trudność, ponieważ BIC (jak i pozostałe miary pokazane w Rozdziale 2) opierają się jedynie na pewnej aproksymacji. Stąd trudno mówić o wyznaczaniu rozkładu zgodnego z hipotezą zerową. Zanim poruszymy bardziej szczegółową tę kwestię przedstawmy alternatywne użyteczne heurystyki.

Pierwszą z nich jest **stopień oznaki** (ang. weight of evidence - WE) obliczany jako logarytm stosunku szans a posteriori. Jego wartość zdaje sprawę z *przewagi* pierwszego modelu nad drugim. W praktyce wykorzystuje się skalę stworzoną przez Jeffreysa [36, s.432]. W zależności od podstawy logarytmu, wyraża się on w **bit**-ach (podstawa=2), **nat**-ach(podstawa=e) lub **ban**-ach(podstawa=10). Dla bitów, wartości powyżej **1.6** świadczą o znacznie większej

wiarogodności pierwszego modelu (opisanego za pomocą licznika stosunku), powyżej **3.3** dostarczają mocnych przesłanek do odrzucenia drugiego modelu, a powyżej **6.6** przyjęcie pierwszego modelu jest bezdyskusyjne (zob. http://en.wikipedia.org/wiki/Bayes factor).

W literaturze (np. w [10, 25]) możemy spotkać się ze współczynnkiem AWE, który jest akronimem od **A**pproximate **W**eight of **E**vidence i wyraża się przez:

$$AWE = -2 \log B_k \tag{4.12}$$

W tym przypadku B_k oznacza czynnik Bayesa dla porównywania modeli z k skupieniami przeciw $k = 1$ (amorficzność struktury). Im większa wartość AWE, tym większa szansa wybrania modelu z daną liczbą skupień. Autorzy tej koncepcji - Banfield i Raftery [10], na podstawie badań symulacyjnych podkreślają, że ważna jest nie tylko nominalna wartość kryterium, ale jego dynamika w zależności od liczby skupień. Dokładniej, optymalnej liczby skupień powinniśmy szukać tam, gdzie następuje spadek tempa zmiany AWE.

Kryterium to było stosowane na początku rozwoju metodologii modeli mieszanych, jednak jak podaje Dasgupta [19, s.298] miara ta stopniowo wychodzi z użycia od momentu wynalezienia algorytmu EM, dzięki któremu możliwe stało się szybkie obliczanie wiarogodności dla modeli mieszanych i korzystanie bezpośrednio z BIC.

4.4. Problemy z testowaniem

Niewątpliwie BIC oraz inne wymienione kryteria są przydatne do porównywania estymowanych modeli. Chcielibyśmy jednak posłużyć się pewnym narzędziem statystycznym. Nasuwa się jednak pytanie, czy w analogiczny sposób możemy wyznaczyć najlepsze modele posługując się klasycznym ilorazem wiarogodności i pochodną statystyką chi-kwadrat?

Pytanie to zostało postawione przez statystyków dość wcześnie. Zajmowali się oni przede wszystkim modelami będącymi mieszaninami rozkładów normalnych. Poniższe problemy dotyczyć więc będą przede wszystkim testowania mieszanek rozkładów normalnych.

4.4.1. Funkcja wiarogodności

Okazuje się, że odpowiedź na zadane wcześniej pytanie jest negatywna. W wielu miejscach ([3, 23, 48]) możemy spotkać się ze stwierdzeniem, że jest rzeczą ogólnie znaną, że nie są spełnione odpowiednie warunki regularności dla rozkładu statystyki $-2 \log \Lambda$ co powoduje, że asymptotycznym rozkładem niekoniecznie jest rozkład ch-kwadrat o liczbie stopni swobody równej róznicy parametrów.

Najczęściej przytaczanym argumentem jest wyjaśnienie znajdujące się w pracy Aitkina i Rubina [3, s.70], którzy uważają, że źródłem problemów jest fakt, że jeśli estymator największej wiarogodności θ znajduje się w otoczeniu granicy przedziału parametrów tj. przyjmuje wartości bliskie 0 lub 1, wówczas rozkład funkcji wiarogodności nie może być opisana za pomocą rozkładu normalnego. Czy taka sytuacja często ma miejsce? Niestety tak, ponieważ przeważająca większość testów jest postaci, że co najmniej jeden z parametrów proporcji λ jest równy 0. Odpowiada to hipotezie na temat braku jednego ze składników mieszaniny. Wówczas testy oparte na asympotycznym rozkładzie chi-kwadrat mogą być zwodnicze.

4.4.2. Zagnieżdżanie

Osobnym problemem jest kwestia zagnieżdżenia modeli (ang. nested models). Jak możemy przeczytać u Silveya [55], test ilorazu wiarogodności bardzo dobrze sprawdza się w

sytuacjach, gdy porównujemy zakładany model z pewnym **modelem nasysconym**. Przypomnijmy, że model nazywamy nasyconym, gdy do wyznaczenia łącznego rozkładu wskaźników wykorzystuje całą informację na temat tych wskaźników. Taki model idealnie spełnia warunek dobroci dopasowania do danych, jest jednak niepraktyczny z uwagi na wysoką maksymalną możliwą liczbę parametrów.

Klasycznym przykładem jest dwuwymiarowy rozkład zmiennych zero-jedynkowych:

Tabela 4.1

X, Y	0	1	
0	π_{00}^{XY}	π_{01}^{XY}	π_0^X
1	π_{10}^{XY}	π_{11}^{XY}	π_1^X
	π_0^Y	π_1^Y	1

Przykładowa hipoteza zerowa głosi, że zmienne są niezależne stochastycznie, a konkurencyjna, że tak nie jest. Zgodnie z pierwszym modelem, łączny rozkład wskaźników jest w całości wyznaczony przez rozkład brzegowy, czyli przez dwa parametry: π_0^Y oraz π_0^X. W przypadku konkurencyjnym mamy model nasycony, który zakłada, że łączny rozkład wskaźników jest wielomianowy z prawdopodobieństwami, których estymatorami są częstości w próbie: π_{00}^{XY}, π_{01}^{XY} i π_{11}^{XY}. Operacja ustalania jednego z parametrów modelu nazywa się **zagnieżdżaniem** go w modelu bardziej skomplikowanym tj. o większej liczbie parametrów.

Okazuje się, że modele można zagnieżdżać nie tylko w modelu nasyconym, ale również porównując je sobą. Spełnione musi być jednak założenie, że jeden z nich jest prostszą wersją drugiego. Możliwe jest wówczas stosowanie testu ilorazu wiarogodności opartego na statystyce G^2 i konstruowanie obszaru krytycznego dla asymptotycznego rozkładu χ^2 o liczbie stopni swobody równej różnicy w liczbie parametrów obu modeli.

Przypadek modeli mieszanych jest jednak nieco bardziej złożony. Powodem dla których Aitkin i Rubin [3] nie byli w stanie przeprowadzić **LRT** dla pośrednich liczb skupień (np. $k = 2$ vs $k = 5$) jest fakt, że na ogół model dwuskładnikowy nie jest szczególnym przypadkiem modelu pięcio-składnikowego. Z postaci modelu nie wynika bowiem, żeby którekolwiek ze skupień w prostszym modelu składało się w oczywisty sposób z dwóch lub więcej składników modelu złożonego.

Jednym przypadkiem, gdy mamy zagwarantowane zagnieżdżanie modeli jest sytuacja, gdy hipoteza zerowa głosi, że w zbiorze nie ma żadnej struktury. Jednak, tak jak pisaliśmy z powodu braku spełnienia warunków regularności hipotetyczny rozkład **LRT** nie jest znany.

4.4.3. Lokalna optymalność rozwiązań

Mieszaniny rozkładów znane są z tego że posiadają wiele (a zatem lokalnych) maksimów funkcji wiarogodności, co oznacza, że model o ustalonej liczbie skupień nie jest wyznaczony jednoznacznie. Mogą istnieć dwie lub więcej parametryzacji modelu, które gwarantują równie dobre dopasowanie do danych. Przyczyna ta może mieć kilka źródeł. Pierwszą z nich jest zbyt duża liczba parametrów koniecznych do oszacowania. Drugą przyczyną jest sama natura algorytmu, który nie przegląda wszystkich rozwiązań, lecz optymalizuje je na zbiorze określonym w poprzednim kroku. Tę niepożądaną właściwość ma m.in. praktycznie jedyny efektywny algorytm do tego typu zagadnień, czyli EM.

4.5. Próby rozwiązania problemów z LRT

Problem braku regularności próbowano rozwiązywać na wiele sposobów. Pierwsze próby opierały się o modyfikację (przeskalowanie) wyjściowej statystyki, jednak wraz z postępem w rozwoju komputerów zaczęto posługiwać się symulacjami (przede wszystkim metodą Monte Carlo). Przełomowym momentem było ujęcie metod symulacyjnych w ramy teoretyczne przez B.Efrona, który nadał im nazwę **bootstrap**, o której wspomnieliśmy już przy okazji wyznaczania liczby klas przy pomocy przedziałów ufności.

Do pierwszych prób można zaliczyć pracę Wolfe [62], który za pomocą symulacji przetestował różne przeskalowania oryginalnej statystyki Λ. Jedną z pierwszych obserwacji był asymptotyczny rozkład χ^2 dla statystyki $\frac{n-3}{n}G^2$ wykorzystanej do testowania $k = 1$ przeciw $k = 2$. Na podstawie tego faktu i dalszych symulacji doszedł on do wniosku, że statystyka $\tilde{\Lambda}$:

$$\tilde{\Lambda} = -\left(\frac{2}{n})(n - p - 2\right) \log \Lambda \tag{4.13}$$

ma rozkład χ^2 o $2p$ stopniach swobody, gdzie p oznacza wymiar przestrzeni. Wolfe [62] uogólnił ten wzór na przypadek testowania k przeciw $k + 1$:

$$\tilde{\Lambda} = -\left(\frac{2}{n})(n - p - 1 - \frac{1}{2}(k + 1)\right) \log \Lambda$$

W tym przypadku $\tilde{\Lambda}$ miałby mieć rozkład χ^2 o $2v - 2$ stopniach swobody, gdzie v jest różnicą w liczbie parametrów obu modeli. Ta sugestia stała się obiektem kontrowersji wśród statystyków. Pierwsze wątpliwości pojawiły się u Hartigana który uznał, że powyższa modyfikacja nie jest konieczna, ponieważ rozkład $-2 \log \Lambda$ *znajduje się gdzieś „pomiędzy”* χ^2_m i χ^2_{p+1}.

Z kolei Everitt [23, s.174] na podstawie symulacji Monte Carlo pokazał prawdziwość zbieżności zasugerowanej przez równanie **??**. Jednak dotyczy to wyłącznie przypadków, gdy liczebność próby n nie przekracza $10p$.

W tym samym roku Aitkin i in. [2] weryfikując wyniki badań nad stylami nauczania za pomocą analizy klas ukrytych zasygnalizowali znany już wówczas problem z ilorazem wiarogodności. Ich propozycja była na ówczesne realia dość nowatorska, lecz z dzisiejszej perspektywy mało precyzyjna. Idea Aitkina i współpracowników opierała się 19-powtórzeniowym bootstrapie tj. symulacji wartości $-2 \log \Lambda$ pod warunkiem braku struktury. Na tej podstawie szacowano teoretyczny rozkład statystyki. W kolejnym kroku liczono dodatkową, dwudziestą wartość statystyki, jednak już na podstawie całej próby. W ten sposób dla 20 osobowej ustalano poziom istotności $\alpha = 0,05$ i następującą regułę decyzyjną: jeśli dodatkowa wartość znajdowała się w obszarze krytycznym tj. istotnie różniła się od wcześniejszych wyników, wówczas odrzucano hipotezę zerową. Problem zagnieżdżania modeli został jednak w tej pracy całkowicie pomięty, gdyż autorzy zaproponowali dokładnie to samo postępowanie w przypadku testowania k przeciw $k + 1$, dla $k > 1$.

Bezpośrednie odwołanie do tej pracy można znaleźć w [48], który co prawda ogranicza się do najprostszej formy testów ($k = 1$, przeciw $k = 2$), ale przeprowadza symulacje w znacznie szerszym zakresie niż czynili to Aitkin i in. [2] Druga różnica polega na dodatkowym elemencie testu, jakim jest założenie o nierówności macierzy kowariancji poszczególnych składowych (ang. heteroscedascity). Okazuje się, bowiem, że o ile w sytuacji równości wariancji przybliżenie 4.13 rzeczywiście daje dobre rezultaty, to w pozostałych przypadkach obserwuje się znaczne odchylenie. Asymptotycznym rozkładem, który najlepiej przybliżał rozkład $-2 \log \Lambda$

pod warunkiem prawdziwości hipotezy o braku struktury w przypadku jednowymiarowego rozkładu normalnego był χ^2_6.

W pracy Aitkin i Rubina [3], w której autorzy dokładnie wyjaśniają przyczyny braku zbieżności $-2\log\Lambda$, poddane są krytyce wcześniejsze obliczenia Wolfe'a. Konstruktywnym elementem tej krytyki jest oryginalna koncepcja odbiegająca od metod symulacyjnych i posiadająca ciekawe uzasadnienie teoretyczne.

Idea autorów polega na obserwacji, że estymacja modelu mieszaniny sprowadza się do oszacowania niezależnych grup parametrów. Do pierwszej grupy należą parametry typu λ odpowiadające za przynależność do danej klasy lub wyznaczające proporcje mieszaniny. Druga grupa składa się z parametrów θ wyznaczających łączny rozkład wskaźników w klasach. Sposób postępowania może być ujęty w następujące kroki:

1. Wybierz pewien rozkład a priori dla parametrów λ. Zauważmy, że przypomina to pierwszy krok algorytmu EM, w którym inicjowane są pewne wartości estymatorów. Główna różnica polega na tym, że teraz rozkład nie jest wyspecyfikowany, lecz przeciwnie - bierzemy pod uwagę wszystkie możliwe sytuacje. Przypomina to wnioskowanie Bayesowskie, gdzie brzegowy rozkład parametru θ jest obliczany jako suma iloczynów prawdopodobieństw warunkowych $p(\theta|\lambda_i)$ i prawdopodobieństw a priori $p(\lambda_i)$. Jak pamiętamy, w ogólnym przypadku, zamiast sumy pojawiała się całka względem miary, która to miara oznaczała rozkład λ.

   Postać tej całki wyraża się wzorem:

$$f(x) = \int \sum_{j=1}^{G} \lambda_j f(x_i|\theta_j), dP(\lambda) \tag{4.14}$$

2. Oblicz maksimum funkcji wiarogodności i wartości estymatorów, które je maksymalizują

Funkcja wiarogodności odpowiadająca powyższej funkcji gęstości dla realizacji x wyraża się poprzez:

$$L(\theta|x) = \int \prod_{i=1}^{n} (\sum_{j=1}^{G} \lambda_j f(x_i|\theta_j)) dP(\lambda) \tag{4.15}$$

Na całkowanie względem miary można patrzeć jak na obliczanie oczekiwanego pola prostokąta, gdy jeden z boków wyraża się poprzez zmienną losową o danym rozkładzie. Mamy zatem do czynienia nie z jednym konkretnym polem, ale z całą rodziną pól na których określony jest rozkład prawdopodobieństwa. Uśrednienie wszystkich wyników jest właśnie całkowaniem względem miary. W przypadku, gdy jeden z boków (dla ustalenia uwagi, nazwijmy go λ) ma rozkład dyskretny, wówczas przestrzeń możliwych wyników jest skończona i w miejsce całki pojawia się znak sumowania po elementach zbioru λ. Analogia jest tym bardziej uzasadniona, że zgodnie z wcześniejszą uwagą dotyczącą działania algorytmu EM, można pokazać ortogonalność wektorów λ i θ.

Maksymalizacja powyższego wyrażenia lub jego logarytmu jest złożona obliczeniowo (por. Rozdział 3) z powodu obecności całki i sumy pod logarytmem. Aby ominąć ten problem, postępujemy podobnie - postać mieszaniny (mixture likelihood) modyfikuje się poprzez dodanie nowej zmiennej oznaczającej klasyfikację (classification likelihood).

Kolejne kroki wykonuje za nas algorytm EM. Dodatkową przeszkodą jest potrzeba numerycznego całkowania w kroku E, co jest pewną niedogodnością. Jak się jednak okazuje, wartości θ, które maksymalizują **??** są często bliskie wartościom estymatorów, które maksymalizują funkcję wiarogodności w tradycyjnym zapisie.

Jednym z najnowszych pomysłów usprawnienia testu jest koncepcja autorstwa Y.Lo, N.Mendell i D.Rubina [46]. Opiera się na **LRT** bazującym na tzw. **kryterium informacyjnym Kullbacka-Leiblera**. Zanim je dokładnie zdefiniujemy, wprowadźmy pewne założenia i oznaczenia (zob. [46, ss.768-770]).

Niech $x_1, x_2, \cdots, x_n$ będzie próbą losową pochodzącą z mieszaniny rozkładów normalnych o gęstości:

$$h(x, \beta) = \sum_{j=1} k\lambda_j \cdot f_i(x, \mu_j, \Sigma_j)$$

Gdzie $\beta = (\lambda_1, \mu_1, \Sigma_1, \cdots, \lambda_k, \mu_k, \Sigma_k)$

Załóżmy teraz, że chcemy zweryfikować hipotezę H_0 głoszącą, że rozkład jest generowany przez k_0 składników, przeciw hipotezie, że tych składników jest k_1, przy czym $k_0 < k_1$. Formalnie, treść obu hipotez można zapisać jako:

$$F_\theta \equiv \{F(x, \theta), \theta \in \Theta \subset R_p\}$$
$$F_\gamma \equiv \{F(x, \gamma), \gamma \in \Gamma \subset R_q\}$$

Gdzie θ i γ są odpowiednio wektorami o podobnej strukturze, co β, przy czym pierwszy z nich ma długość $p = 3k_0 - 1$, a drugi $1 = 3k_1 - 1$. W ten sposób model zgodny z H_0 jest zagnieżdżony w H_1 ponieważ ten pierwszy można uzyskać poprzez wyzerowanie lub ustalenie niektórych parametrów modelu H_1.

Kryterium informacyjne Kullbacka-Leiblera definiujemy następująco:

$$I(h : f : \theta) = E_h \log \frac{h(x, \beta)}{f(x, \theta)}$$
$$I(h : g : \gamma) = E_h \log \frac{h(x, \beta)}{g(x, \gamma)}$$

Gdzie f i g oznaczają gęstości pojedynczych składników mieszaniny rozkładów. Uznaje się, że model H_1 lepiej przybliża empiryczny rozkład wskaźników niż model H_0 wtedy i tylko wtedy, gdy:

$$\sup_\theta E_h\{\log f(x, \theta)\} > \sup_\gamma E_h\{\log g(x, \gamma)\}$$

W oryginalnym tekście [46, s.770-771] można znaleźć dokładny opis warunków regularności, dla których zachodzą odpowiednie twierdzenia graniczne. Nas interesować będzie tylko końcowy wniosek:

$$LR = \sum_{i=1}^{n} \log \frac{f(x_i, \hat{\theta})}{g(x_j, \hat{\gamma})}$$

Gdzie $\hat{\theta}, \hat{\gamma}$ są estymatorami największej wiarogodności swoich parametrów. W oparciu o tak zdefiniowaną statystykę testową określa się parę następujących hipotez. Pierwsza z nich mówi, że oba modele o k_0 i k_1 składnikach jednakowo dobrze przybliżają empiryczny rozkład wskaźników tj.

$$E_h\{\log f(x, \hat{\theta})\} = E_h\{\log g(x, \hat{\gamma})\}$$

Druga zaś, że lepsze przybliżenie daje model o większej liczbie parametrów.

$$E_h\{\log f(x, \hat{\theta})\} > E_h\{\log g(x, \hat{\gamma})\}$$

Najbardziej istotną rzeczą jest określenie granicznego rozkładu statystyki testowej. W [46, s.772] można znaleźć odpowiednie twierdzenie, które mówi, że dystrybuanta statystyki $2LR$ dąży według rozkładu do dystrybuanty zmiennej będącej sumą $p+q$ zmiennych o rozkładzie χ^2_1. Wagami tej kombinacji są wartości własne macierzy informacyjnej zbudowanej z kolejnych pochodnych cząstkowych kryterium Kullbacka-Leiblera. Dokładna postać tej macierzy wraz z dowodem znajduje się w [46, s.772].

4.6. Podsumowanie

Możliwość testowania liczby klas może być przełomowym momentem dla rozwoju metodologii analizy skupień. Jest tak, ponieważ dzięki Wykorzystaniu procedur statystyki inferencyjnej dysponujemy informacją, dzięki której możemy uzasadnić decyzje na temat liczby klas. Nadal jednak przeprowadzenie odpowiednich testów ma mocno ograniczony charakter. Większość opisanych procedur odnosi się do szczególnego przypadku rozkładów normalnych oraz testowania hipotezy zerowej na temat braku struktury. Zanim jednak nie zostaną rozstrzygnięte podstawowe problemy dla tego typu uproszczonych sytuacji, nie ma sensu rozpatrywać przypadków bardziej złożonych.

Rozważania na temat możliwości testowania hipotez na temat liczby klas ujawniają specyfikę analizę skupień względem innych analiz statystycznych. Okazuje się, że w wielu przypadkach standardowe metody konstruowania testów nie mają zastosowania. Wybór konkretnej liczby skupień nie jest tożsamy z wyborem odpowiadającego jej konkretnego modelu. Model o ustalonej liczbie skupień nie jest wyznaczony jednoznacznie. Mogą istnieć dwie lub więcej parametryzacji modelu, które gwarantują równie dobre dopasowanie do danych.

Ponadto, bezrefleksyjne kopiowanie metod z innych obszarów wnioskowania może być zwodnicze. Dla typowego w podobnych sytuacjach testu ilorazu wiarogodności nie są spełnione odpowiednie warunki regularności. Najczęstszym problemem jest fakt, że testowana wartość jednego z parametrów leży w bliskim otoczeniu zera, a więc na granicy przestrzeni parametrów. W ten sposób asyptotyczny rozkład statystyki $-2\log\Lambda$ niekoniecznie musi być rozkładem chi-kwadrat o liczbnie stopni swobody równej różnicy parametrów między testowanymi modelami.

Osobnym problemem jest kwestia zagnieżdżenia modeli, zwłaszcza w przypadku weryfikacji liczby klas, gdy hipoteza zerowa jest inna niż ta o braku struktury. Trudno bowiem orzec, w jakim stopninu model z większą liczbą klas jest szczególnym przypadkiem modelu mniej skrystalizowanego.

Do najbardziej znanych prób rozwiązania problemów z testowaniem zalicza się: modyfikacje statystyki $-2\log\Lambda$, próbkowanie metodą bootstrap, całkowanie względem rozkładu jednej grupy parametrów oraz wykorzystanie kryterium informacyjnego Kullbacka-Leiblera.

W pakietach statystycznych, do wyboru optymalnej liczby skupień najczęściej wykorzystuje się kryteria oparte na czynniku bayesowskim - AWE lub na wartości zmaksymalizowanej funkcji wiarogodności - BIC.

Dokładniej o możliwościach wybranych pakietów powiemy w następnym rozdziale.

Rozdział 5

Analiza klasy ukrytej i modele mieszane w praktyce

5.1. Opis symulacji

Wcześniejsze rozważania miały na celu pokazanie podstawowych problemów teoretycznych analizy skupień. Ze względu na szeroką gamę dostępnych metod, porównanie wszystkich ze wszystkimi znacznie przekroczyłoby ramy tej pracy. Ponadto, w ubiegłych latach wykonano wiele podobnych owocnych analiz (dla przykładu zob. [**?**, 50] i chcielibyśmy uniknąć ewentualnego powielenia cudzych wniosków.

Dlatego zdecydowaliśmy się ograniczyć do wybranej klasy metod i algorytmów. Ich wspólną cechą jest modelowe ujęcie analizy skupień i potraktowanie jej w kategoriach probabilistycznych. Skupimy się na następujących pakietach: **SPSS 16**, **R** oraz **LatentGold 4.0**.

W pierwszym z nich wykorzystamy algorytm **grupowania dwustopniowego** (ang. Two-Step), który według twórców pakietu wykorzystuje metody statystyczne przy grupowaniu obiektów. Krótki, ale wyczerpujący opis algorytmu można znaleźć w Aneksie tej pracy lub w raporcie technicznym SPSS [57].

W odniesieniu do kolejnego pakietu powinniśmy raczej używać słowa „środowisko", ponieważ **R** jest nieustannie aktualizowany przez użytkowników poprzez programowanie nowych skryptów i pakietów. Jednym z nich jest jest pakiet **MCLUST** (od ang. Model-Based Clustering), którego twórcami są znani nam z poprzednich rozważań C. Fraley i A. Raftery.

Przyjrzymy się również pakietowi stworzonemu na potrzeby analizy skupień za pomocą analizy klasy ukrytej. Jego autorzy również pojawili się we wcześniejszych cytowaniach, a są nimi J.Vermunt i J.Magidson).

Mamy nadzieję pokazać omawiane zalety metod probabilistycznych, jeśli chodzi o dostarczanie danych do podjęcia uzasadnionych decyzji. Celem poniższego zestawienia jest przede wszystkim porównanie możliwości wybranych pakietów, niż ich systematyczna krytyka. Niemniej jednak istnieje potrzeba wykonania takich badań dla wszystkich szerzej wykorzystywanych algorytmów modelowej analizy skupień.

5.2. Opis danych wejściowych

Przeglądu metod dokonamy w oparciu o **trzy** różne zbiory danych. Pierwszy z nich już częściowo poznaliśmy i jest nim zbiór o skrystalizowanej strukturze - zbiór $\mathcal{S}$. Drugim z nich jest zbiór amorficzny oznaczony jako $\mathcal{A}$. Przypomnijmy, że są to zbiory dwuwymiarowe o ciągłych wskaźnikach. Pierwszy z nich powstał z myślą badania **skuteczności** w wykrywaniu

istniejącej struktury dziewięciu skupień, podczas gdy drugi ma na celu ocenę **wrażliwości** algorytmu na brak jakiejkolwiek struktury w zbiorze.

Pierwsze dwa zbiory posłużą nam do porównania różnych metod zajmujących się identyfikacją parametrów rozkładów normalnych w modelu mieszanym. Porównamy następujące pakiety: **SPSS TwoStep z metodą odległości euklidesowej**, pakiet **R MCLUST** oraz **Latent Gold**.

Trzecim zbiorem, o którym jeszcze nie mówiliśmy jest zbiór **Latent**, który jest opisany za pomocą dwóch klas ukrytych i czterech wskaźników o charakterze dychotomicznym. Zostanie on wykorzystany się do weryfikacji poprawności analizy skupień w oparciu o model klasy ukrytej. W tym miejscu porównamy pakiety **TwoStep** z metryką logarytmu wiarogodności oraz **Latent Gold**.

5.3. Rzeczywiste wartości parametrów

Na wstępie udostępnimy informacje na temat prawdziwych wartości parametrów, na podstawie których zostały wygenerowane poniższe zbiory. Dane zawarte w tabelach mają na celu służyć jako punkt odniesienia do uzyskanych rezultatów za pomocą każdej z metod.

5.3.1. Zbiór amorficzny

Zbiór $\mathcal{A}$ ze względu na swoją prostą strukturę jest opisany za pomocą dwóch identycznych par parametrów: $\mu_x = \mu_y = 0$ oraz $\sigma_x = \sigma_y = 1$. Liczebność zbiory wynosi $n = 300$.

5.3.2. Zbiór skrystalizowany

Tabela 5.1: Parametry modelu skrystalizowanego, k=9

nr skupienia	1	2	3	4	5	6	7	8	9	out
Częstość skupień	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
μ_X	40	15	20	5	40	25	30	10	35	15
σ_X	1	1	1	1	1	1	1	1	1	75
μ_Y	5	5	25	5	10	20	25	10	5	15
σ_Y	1	1	1	1	1	1	1	1	1	75

Podobnie jak zbiór amorficzny, liczy sobie $n = 300$, z czego 270 posiada wyraźną klasyfikację, natomiast pozostała część została wygenerowana z rozkładu jednostajnego jako jednostki odstające.

5.3.3. Zbiór z klasą ukrytą

Tabela 5.2: Zbiór Latent: Parametry modelu z klasą ukrytą

	X	A	B	C	D
1	0.5	0.8	0.9	0.7	0.7
2	0.5	0.4	0.2	0.3	0.3

W efekcie otrzymaliśmy zbiór o liczebności $n = 1000$ o rozkładzie profili opisanym za pomocą tabel 5.2 oraz 5.3.

Tabela 5.3: Zbiór Latent: Rozkład liczebności profili

profil	liczebność
"1111"	180
"1110"	84
"1101"	84
"1011"	34
"0111"	50
"1100"	52
"1010"	42
"0110"	31
"1001"	42
"0101"	31
"0011"	27
"1000"	82
"0100"	38
"0010"	52
"0001"	52
"0000"	119

5.4. Wyniki i wnioski

Otrzymane wyniki będziemy porównywać pod kątem cech „dobrej" metody analizy skupień, o której mówiliśmy w Rozdziale 2. Przypomnijmy, że aspekty, o których będzie mowa stanowią następujący zbiór zagadnień:

1. Segmentowalność zbioru i liczba skupień
2. Optymalność podziału
3. Jednostki odstające

5.4.1. Segmentowalność zbioru i liczba skupień

Model dyskretny

W przypadku metody TwoStep decyzja podejmowana jest automatycznie w oparciu o wartość BIC. W naszym przypadku metoda automatycznego wyboru liczby klas oraz narzuconego przez użytkownika doprowadziła do identycznych rezultatów. Mimo, że najniższa wartość kryterium osiągana jest dla czterech klas, to największa zmiana kryterium nastąpiła „między" $k = 1$ i $k = 2$, co uzasadniło wybór takiej liczby klas.

Tabela 5.4: TwoStep: Wartości BIC dla różnej liczby skupień

Liczba skupień	BIC
1	5522.521
2	4314.952
3	3626.676
4	2978.148

Znacznie większą wartość BIC uzyskujemy dla modelu estymowanego za pomocą pakietu Latent Gold. W tym miejscu jest ona równa 5287. Została ona obliczona na podstawie wartości funkcji wiarogodności (w tym miejscu -2612). Dodatkową informacją jest przybliżona oznaka istnienia struktury w zbiorze (AWE=5982).

Tabela 5.5: Latent Gold: Podsumowanie modelu

Logarytm funkcji wiarogodności	-2612.62
BIC	5287.418
AWE	5982.727
Wartość statystyki X^2	0.0476
Liczba stopni swobody	6
p-value	1

Model ciągły

Zbiór skrystalizowany Analogicznie do przypadku dyskretnego, wybierany jest model o najniższej wartości BIC. Automatyczny wybór liczby skupień wskazuje na obecność struktury o trzech klasach (BIC = 2998) - zob. tabela **??**.

Tabela 5.6: TwoStep: wybór wartości dla różnej liczby skupień

Liczba skupień	BIC
1	3278.32
2	3119.69
3	**2998.96**
4	3010.42
5	3024.84
6	3045.67
7	3065.19
8	3085.91
9	3107.42
10	3129.74
11	3143.66
12	3158.68
13	3172.76
14	3187.78
15	3210.27

Ze względu na brak globalnej optymalności rozwiązania modelu mieszanego za pomocą algorytmu EM w pakiecie **Latent Gold**, postanowiliśmy przeprowadzić 20 estymacji wybierając różne punkty startowe dla wyjściowych parametrów. Spośród lokalnie optymalnych rozwiązań wybraliśmy te o największej wartości funkcji wiarogodności. Podobnie jak w przypadku dyskretnym, kolejnymi informacjami są: wartość BIC liczona w oparciu o funkcje wiarogodności oraz AWE.

Pakiet **MCLUST** zamiast pojedynczej informacji na temat zależności wartości BIC od liczby skupień, dostarcza informację poszerzoną o typ modelu. Zauważmy, że wybór modelu może prowadzić do dużych rozbieżności, jeśli chodzi o optymalną liczbę skupień. Przykładowo, dla modelu „VII" najlepiej wydzielić dziewięć, podczas gdy dla „VEV" tylko trzy skupienia.

Tabela 5.7: Latent Gold: Podsumowanie modelu

logarytm funkcji wiarogodności	-1877.38
BIC	4005.73
AWE	4457.53

Tabela 5.8: MCLUST: Wartości BIC w zależności od typu modelu i liczby skupień

	typ modelu									
k	EII	**VII**	EEI	VEI	EVI	VVI	EEE	EEV	VEV	VVV
1	4612.5	4612.5	4564.1	4564.1	4564.1	4564.1	4569.7	4569.7	4569.7	4569.7
2	4535.3	4420.8	4533.9	4413.3	4539.5	4415.1	4409.1	4395.9	4312.8	4306.5
3	4211.6	4205.6	4136.1	4133.0	4134.4	4133.2	**4141.5**	4146.1	4142.5	4145.6
4	4226.2	4154.2	4153.3	4112.5	4155.0	4103.0	4158.6	4160.0	4090.5	4084.2
5	4242.2	4142.6	4170.4	4112.0	4175.0	4105.5	4175.7	4156.6	4057.9	4057.2
6	4206.0	4118.5	4135.1	4097.8	4127.2	4100.1	4161.4	4082.8	4049.3	4050.9
7	**4185**	4064.2	4133.2	4066.5	**4107.9**	4087.8	4170.22	4084.5	3989.9	3990.5
8	4202.0	4007.7	**4124.3**	4009.9	4165.2	4037.6	4186.8	4053.2	3994.1	3990.9
9	4206.1	**3940.3**	4162.4	**3942.1**	4171.2	**3973.5**	4165.2	**4035.3**	**3971.4**	**3986.2**
min.	4185.0	**3940.3**	4124.3	3942.1	4107.8	3973.5	4141.5	4035.2	3971.3	3986.1

Globalnie, największa wartość kryterium przyjmowana jest dla modelu „VII", co oznacza, że wybrany został model o lokalnie niezależnych wskaźnikach oraz o różnych liczebnościach skupień (zob. Tabela **??**).

Tabela 5.9: MCLUST: Podsumowanie modelu

typ modelu	VII
estymowana liczba skupień	9
logarytm funkcji wiarogodności	-1870.328
BIC	3940.3

Zbiór amorficzny Metoda **TwoStep** z automatycznym wyborem liczby skupień **nie wykazała wrażliwości** na brak struktury decydując się ostatecznie na podział na 3 skupienia. O braku sensowności takiej decyzji możemy przekonać się obserwując ilustrację uzyskanego podziału na przykładzie wykresu rozrzutu oraz indeksu sylwetki. (tu wstawić oba rysunki)

W przypadku **Latent Gold** do testowania braku struktury postanowiliśmy wykorzystać metodę przybliżonego testu ilorazu wiarogodności. Przeprowadziliśmy kolejno estymację dla $k = 1$ oraz $k = 2$. Następnie przy użyciu metody bootstrap uzyskaliśmy 500 wartości statystyki $-2 \log \Lambda$ i graficznie (na tzw. wykresie „QQ-plot") przedstawiliśmy jej rozkład względem hipotetycznych rozkładów chi-kwadrat. Najlepszą aproksymacje uzyskaliśmy dla rozkładu chi-kwadrat o 7 stopniach swobody. (tu rysunek)

Empiryczna wartość statystyki $-2 \log \Lambda$ dla całej próby wyniosła $12,001$, której odpowiada *p-value* równe 0.1 co przemawia za utrzymaniem hipotezy zerowej na temat braku struktury w zbiorze.

Estymacja za pomocą algorytmu EM w pakiecie **MCLUST** dla każdego typu modelu osiągnęła maksymalną wartość BIC dla jednej klasy ukrytej, co potwierdza poprawną identyfikację braku struktury w zbiorze. Natomiast jeśli chodzi o łączny rozkład wskaźników,

Tabela 5.10: Latent Gold: Podsumowanie modelu amorficznego

logarytm funkcji wiarogodności	-1066.8
BIC	2156.42
AWE	2191.23

globalnie optymalny wynik został osiągnięty dla modeli „EII" oraz „VII" (BIC=-2151.5) co w przypadku jednej klasy ukrytej sprowadza się do wyboru tego samego typu modelu mieszanego (oznaczać będziemy go jako „XII").

Tabela 5.11: MCLUST: Wartości BIC w zależności od typu modelu

	typ modelu									
k	EII	VII	EEI	VEI	EVI	VVI	EEE	EEV	VEV	VVV
1	**2151.5**	**2151.5**	2156.4	2156.4	2156.4	2156.4	2160.4	2160.4	2160.4	2160.4

Tabela 5.12: MCLUST: Podsumowanie modelu amorficznego

typ modelu	„XII"
estymowana liczba skupień	1
logarytm funkcji wiarogodności	-1067.2
BIC	2151.51

5.4.2. Optymalność podziału

Model dyskretny

Trudno jest mówić o jakiejkolwiek mierze optymalności podziału w przypadku metody **TwoStep** (zarówno w przypadku dyskretnym jak i ciągłym), ponieważ model, na którym opiera się metoda jest dość złożony. O ile drzewo uzyskane w pierwszym kroku algorytmu można opisać w kategoriach prawdopodobieństwa, o tyle drugi etap polega na grupowaniu hierarchicznym, dla którego trudno znaleźć odpowiedni model skalowania. Podział uzyskany za pomocą grupowania dwustopniowego jest w gruncie rzeczy deterministyczny tj. każda obserwacja należy do danego skupienia z prawdopodobieństwem równym 0 lub 1.

Zupełnie inny wynik otrzymujemy wykorzystując pakiet **Latent Gold**. Warto podkreślić, że o ile w przypadku wielu klasycznych metod analizy skupień faktycznie „inny nie znaczy lepszy", o tyle w tej sytuacji dysponujemy jasnym kryterium optymalności podziału, jakim jest statystyka X^2. Jej wartość wraz z liczbą swobody asymptotycznego rozkłady chi-kwadrat odczytujemy z tabeli 5.5. Jest ona równa różnicy między liczbą parametrów w modelu nasyconym (=15) i modelu z dwiema klasami ukrytymi (=9). Widzimy, że otrzymany zestaw estymatorów odtwarza łączny rozkład wskaźników bardzo dokładnie (nie ma podstaw do odrzucenia hipotezy zerowej mówiącej, że dwa rozkłady: wyjściowy i odtwarzany różnią się w sposób istotny statystycznie).

Uzyskany podział znacznie różni się od podziału uzyskanego metodą **TwoStep**. Mianowicie, w wyniku działania algorytmu EM otrzymujemy dwa rodzaje rozwiązań problemu. Pierwszym z nich jest **klasyfikacja probabilistyczna** (lub rozmyta), w której każda obserwacja należy do każdego ze skupień z różnym prawdopodobieństwem. Druga klasyfikacja jest efektem **dyskretyzacji** owego rozmycia.

Tabela 5.13: TwoStep: podział profili

profil	1	2
"1111"	1	0
"1110"	0	1
"1101"	1	0
"1011"	0	1
"0111"	0	1
"1100"	0	1
"1010"	0	1
"0110"	0	1
"1001"	0	1
"0101"	0	1
"0011"	0	1
"1000"	0	1
"0100"	0	1
"0010"	0	1
"0001"	0	1
"0000"	0	1

Tabela 5.14: Latent Gold: Estymowane wartości parametrów w modelu z klasą ukrytą

	Estymatory	
	1	2
X	0.5011	0.4989
A	0.7997	0.3994
B	0.8968	0.2016
C	0.7006	0.2985
D	0.7006	0.2985

Tabela 5.15: Latent Gold: Liczebności profili estymowane i rzeczywiste

	liczebności		
profil	estymowane	rzeczywiste	składniki X^2
"1111"	179.97	180	3.54E-06
"1110"	83.79	84	0.000507
"1101"	83.79	84	0.000507
"1011"	34.47	34	0.00652
"0111"	49.56	50	0.003835
"1100"	51.98	52	5.78E-06
"1010"	41.98	42	3.6E-06
"0110"	31.53	31	0.008911
"1001"	41.98	42	3.6E-06
"0101"	31.53	31	0.008911
"0011"	26.40	27	0.01362
"1000"	81.99	82	2.46E-07
"0100"	37.79	38	0.001109
"0010"	52.26	52	0.001368
"0001"	52.26	52	0.001368
"0000"	118.65	119	0.001
suma	1000	1000	0.047675

Tabela 5.16: Latent Gold: Dwa rodzaje klasyfikacji

	Według modalnej		
Rozmyta (probabilistyczna)	1	2	Suma
1	461.496	39.6237	501.12
2	84.5039	414.376	498.88
Suma	546	454	1000
Błąd dyskretyzacji	0.124		

W procesie klasyfikowania rozmytego dla każdej obserwacji tworzona jest zmienna o częstościach równym stopniom przynależności do danej klasy. Dyskretyzacja polega na redukcji informacji zawartej w rozkładzie do wartości modalnej tego rozkładu. Innymi słowy obserwacja klasyfikowana jest do skupienia o największym prawdopodobieństwie przynależności pod warunkiem określonego profilu. W wyniku dyskretyzacji pojawiają się „błędy zaokrągleń", co prowadzi z jednej strony prowadzi do wyraźnego podziału zbioru, z drugiej jednak tracona jest cenna informacja na temat „rozmieszczenia" obserwacji między skupieniami. Wielkość błędu dyskretyzacji (w naszym przypadku wynosi on 0.124 - również cenną informacją na temat ogólnej niepewności klasyfikacji. Jeśli macierz podziału zawiera elementy bliskie 0 lub 1, wówczas błąd dyskretyzacji jest nieznaczny.

Porównując wartości estymatorów i odpowiadającym im rzeczywistych parametrów możemy zaobserwować wysoką dokładność oszacowania. Zbiorczą informację na temat optymalności podziału zawiera statystyka X^2, której wartość w próbie wynosi 0.476. Zauważmy, że najlepsze oszacowania otrzymaliśmy dla profili o największych liczebnościach w zbiorze, a najmniej dokładne dla najmniej licznych. Podobnie jest w przypadku warunkowych prawdopodobieństw przynależności do skupień - profile występujące najczęściej cechują się najmniejszym poziomem niepewności (entropii) klasyfikacji.

Tabela 5.17: Latent Gold: Estymowane warunkowe prawdopodobieństwa przynależności do klas ukrytych

profil	modalna	1	2
"0000"	2	0.0078	0.9922
"0001"	2	0.0416	0.9584
"0010"	2	0.0416	0.9584
"0011"	2	0.1925	0.8075
"0100"	2	0.2135	0.7865
"0101"	1	0.5988	0.4012
"0110"	1	0.5988	0.4012
"0111"	1	0.8913	0.1087
"1000"	2	0.0452	0.9548
"1001"	2	0.2066	0.7934
"1010"	2	0.2066	0.7934
"1011"	1	0.5887	0.4113
"1100"	1	0.6198	0.3802
"1101"	1	0.8996	0.1004
"1110"	1	0.8996	0.1004
"1111"	1	0.9801	0.0199

Porównując powyższe wyniki mogą nasunąć się dwa wnioski: albo podział na dwie klasy jest artefaktem albo któryś algorytm działa niepoprawnie. Jeśli miałby to być artefakt, to dlaczego algorytm **TwoStep** zdecydował się na dokładnie dwa skupienia, a Latent Gold zwraca wysoki Goodness-of-Fit?. W jakim stopniu możemy porównywać uzyskane wyniki za pomocą metod statystycznych? Kuszącym wydaje się obliczenie estymatorów parametrów dla metody **TwoStep** i zastosowanie wnioskowania w oparciu o model cechy ukrytej jednak w Aneksie można zauważyć, że zaproponowany tam model grupowania istotnie różni się od modelu klasy ukrytej (np. nie jest spełniony kluczowy aksjomat o lokalnej niezależności wskaźników, na podstawie którego mielibyśmy odtworzyć ich łączny rozkład).

Porównując ostateczne, dyskretne klasyfikacje obiektów uzyskane za pomocą metody **Two**

Step i **Latent Gold** możemy zaobserwować istotne różnice. Oszacowania parametrów pochodzące z pierwszej metody są mniej dokładne, co objawia się wysoką wartością statystyki X^2. Z drugiej strony, podział uzyskany za drugiej metody jest bardziej „tolerancyjny" niż w przypadku **TwoStep**. W przypadku tej ostatniej, do pierwszej klasy zostały zaliczone wyłącznie te profile, które miały co najmniej trzy „jedynki", z których dokładnie dwie stały na miejscu zmiennych A i B. Podział za pomocą **Latent Gold** do pierwszej klasy przyporządkowuje obiekty posiadające co najmniej trzy „jedynki" lub dwie „jedynki", ale na miejscu zmiennej B. Jednak w przypadku drugiego rodzaju profili decyzja jest gorzej uzasadniona ze względu na wysoki stopień niepewności.

Model ciągły

Amorficzny Zarówno **Latent Gold** jak i **MCLUST** dają dość dobre i niemal identyczne oszacowania wartości oczekiwanych, ale wyraźnie gorzej wypadają pod względem estymacji wariancji. Większym błędem cechuje się metoda **MCLUST**.

Tabela 5.18: Latent Gold: Podsumowanie modelu

logarytm funkcji wiarogodności	-1877.38
BIC	4005.73
AWE	4457.53

Tabela 5.19: Latent Gold: Estymatory parametrów

$\hat{\mu}_X$	0.099
$\hat{\sigma}_X$	1.74
$\hat{\mu}_Y$	0.085
$\hat{\sigma}_Y$	1.39

Komentarz

Tabela 5.20: MCLUST: Estymatory parametrów

$\hat{\mu}_X$	0.099
$\hat{\sigma}_X$	2.053
$\hat{\mu}_Y$	0.086
$\hat{\sigma}_Y$	2.053

Skrystalizowany Podział uzyskany metodą **TwoStep** jest mniej dokładny niż podziały uzyskane pozostałymi metodami. Należy to rozumieć jako brak rozpoznania mniejszych skupień w obrębie wyróżnionych trzech. Z drugiej strony, uzyskano równomierny rozkład częstości skupień. Co więcej oszacowane wartości oczekiwane dla trzech skupień są dobrym przybliżeniem średniej z wartości oczekiwanych dla dziewięciu skupień.

Analogicznie, otrzymujemy zestawienie klasyfikacji rozmytej i według modalnej oraz błąd wynikający z dyskretyzacji tej pierwszej. W tym przypadku poniesiony koszt wynosi niecałe 3% całego zbioru obserwacji.

Powyższa tabela przedstawia estymowane wartości parametrów w poszczególnych skupieniach tj. średnie i odchylenia standardowe obu zmiennych. Widzimy, że z dokładnością

Tabela 5.21: TwoStep: Rozkład częstości skupień

skupienie	częstość
1	0.31
2	0.31
3	0.31
outlier	0.07

Tabela 5.22: TwoStep: Estymatory parametrów

nr skupienia	$\hat{\mu}_X$	$\hat{\sigma}_X$	$\hat{\mu}_Y$	$\hat{\sigma}_Y$
1	24.93	4.58	23.41	2.83
2	10.16	4.15	6.32	2.73
3	39.58	4.39	6.67	2.85
outlier	28.31	18.13	17.56	9.02

Tabela 5.23: Latent Gold: Dwa rodzaje klasyfikacji

	Według modalnej									
Rozmyta (probabilistyczna)	1	2	3	4	5	6	7	8	9	Suma
1	61.22	0.005	0	0	1.101	0	0	0	0.0409	62.37
2	0	32.50	0	0.058	0	0	0	0.519	0.0602	33.14
3	0	0	30.679	0	0	0.308	0.003	0	0	30.99
4	0	0.007	0	30.435	0	0	0	0.470	0	30.91
5	0.042	0	0	0	30.332	0	0	0	0	30.37
6	0	0	0.196	0	0	29.62	0.023	0	0.36	30.20
7	0	0	0.0004	0	0	0.008	29.09	0	0.0043	29.10
8	0	0.043	0	0.334	0	0	0	28.649	0	29.027
9	0.73	0.43	1.12	0.17	0.56	1.05	0.88	0.36	18.53	23.86
Suma	62	33	32	31	32	31	30	30	19	300
Błąd dyskretyzacji	0.0298									

Tabela 5.24: Latent Gold: Estymatory parametrów

	1	2	3	4	5	6	7	8	9	out
Częstość skupień	0,207	0,110	0,103	0,103	0,101	0,100	0,097	0,096	0,079	0
$\hat{\mu}_X$	39.674	14.979	19.881	5.588	39.722	24.912	30.416	9.831	27.252	-
$\hat{\sigma}_X$	5.400	1.756	1.668	1.709	1.482	1.325	1.468	1.306	18.399	-
$\hat{\mu}_Y$	4.875	4.681	25.202	4.858	10.197	19.931	25.059	9.588	18.683	-
$\hat{\sigma}_Y$	1.343	1.868	1.686	1.437	1.374	1.360	1.251	1.363	7.725	-

Tabela 5.25: Latent Gold: różnice między wartościami estymatorów i prawdziwych parametrów

	1	2	3	4	5	6	7	8	9	out
$\frac{\hat{\mu}_X - \mu_X}{\hat{\sigma}_X}$	0.060	0.012	0.071	0.344	0.187	0.066	0.284	0.129	0.421	-
$\frac{\hat{\mu}_Y - \mu_Y}{\hat{\sigma}_Y}$	0.092	0.170	0.120	0.098	0.144	0.050	0.047	0.302	0.931	-

do permutacji skupień uzyskano dość precyzyjne oszacowania. Wyjątkiem jest skupienie, dla którego estymowana częstość jest dwukrotnie wyższa niż w rzeczywistości.

W przypadku oceny stopnia odtwarzania rozkładu cechy ukrytej (proporcji skupień) możliwe jest wykorzystanie statystyki X^2 na podobnej zasadzie, jaką posłużyliśmy się w poprzednich przykładach. Ta metoda nie sprawdza się jednak w przypadku porównania średnich. Przedstawione poniżej wyniki są ideowo zbliżone do statystyki **t-studenta** wykorzystywanej do testowania hipotez na temat równości średnich w dwóch rozłącznych grupach. Wartość bezwzględną różnicy między średnimi estymowanymi i średnimi rzeczywistymi podzielono za każdym razem przez wartość odchylenia standardowego w próbie.

Tabela 5.26: Latent Gold: Klasyfikacja profili ze względu na zakresy wartości pojedynczych wskaźników

nr skupienia	1	2	3	4	5	6	7	8	9
X									
0.366 - 10.42	0	0.0026	0	0.515	0	0	0	0.3913	0.0911
10.47 - 20.37	0.0001	0.5494	0.3066	0.0002	0	0	0	0.0925	0.0512
20.58 - 30.11	0.0007	0.0004	0.21	0	0	0.5034	0.2105	0	0.0751
30.12 - 39.01	0.517	0	0	0	0.1541	0	0.2746	0	0.0544
39.31 - 49.91	0.5217	0	0	0	0.3522	0	0	0	0.1261
Y									
0.0500 - 4.777	0.4292	0.2798	0	0.2649	0	0	0	0.0003	0.0258
4.815 - 6.566	0.5611	0.2315	0	0.1944	0.0003	0	0	0.0051	0.0076
6.586 - 11.46	0.0491	0.0404	0	0.0559	0.4095	0	0	0.431	0.0141
11.67 - 23.61	0	0.0006	0.0964	0	0.0964	0.5029	0.0461	0.0475	0.21
23.67 - 29.85	0	0	0.4201	0	0	0.0005	0.4389	0	0.1404

5.4.3. Jednostki odstające

Ponieważ zarówno w zbiorze $\mathcal{A}$ oraz **Latent** nie zakładaliśmy istnienia jednostek odstających, w tym miejscu będziemy mówić tylko o modelu ciągłym dla zbioru skrystalizowanego. Okazuje się, że ani **Latent Gold** ani **MCLUST** nie są wrażliwość obecność jednostek odstających. Za każdym razem włączane były do utworzonych skupień. Równoważnie, oznacza to, że były „obejmowane" przez gęstości łącznego rozkładu wskaźników wyznaczonego przez dane skupienie. Często wymuszało to zwiększenie zakresu tolerancji, czyli wariancji rozkładu. To z kolei prowadziło do przeszacowania estymatorów wariancji w skupieniach.

Jeśli chodzi o metodę **TwoStep** to w tabeli **??** pojawia się dodatkowa informacja na temat klasyfikacji jednostek odstających. Ich dokładna klasyfikacja została przedstawiona w tabeli **??**.

W analogiczny sposób, jak zostało to zaprezentowane w Rozdziale 2 na temat możliwych

Tabela 5.27: Latent Gold: Klasyfikacja wybranych profili ze względu na łączne wartości wskaźników

X	Y	modalna	1	2	3	4	5	6	7	8	9
17.17	4.228	2	0.0002	0.9936	0	0	0	0	0	0	0.0062
17.28	25.82	3	0	0	0.9711	0	0	0	0	0	0.0289
17.36	26.04	3	0	0	0.9719	0	0	0	0	0	0.0281
17.37	27.96	3	0	0	0.9335	0	0	0	0	0	0.0665
17.41	6.209	2	0.0003	0.985	0	0	0	0	0	0	0.0147
17.77	11.72	9	0	0.0384	0	0	0	0	0	0	0.9616
17.78	26.18	3	0	0	0.9778	0	0	0	0	0	0.0222
18.21	23.52	3	0	0	0.9686	0	0	0	0	0	0.0314
18.22	21.84	3	0	0	0.8689	0	0	0.0007	0	0	0.1304
18.64	24.31	3	0	0	0.9822	0	0	0	0	0	0.0178
18.67	25.73	3	0	0	0.986	0	0	0	0	0	0.014
18.94	3.049	2	0.0026	0.9688	0	0	0	0	0	0	0.0286
19.06	24.9	3	0	0	0.9866	0	0	0	0	0	0.0134
19.12	23.59	3	0	0	0.9773	0	0	0.0001	0	0	0.0226
19.37	26.42	3	0	0	0.9864	0	0	0	0	0	0.0136
19.48	1.124	2	0.0018	0.8863	0	0	0	0	0	0	0.1119

Tabela 5.28: MCLUST: Estymatory parametrów

	1	2	3	4	5	6	7	8	9	out
Częstość skupień	0.082	0.228	0.084	0.103	0.102	0.102	0.103	0.100	0.096	0.1
$\hat{\mu}_X$	5.16	12.08	27.54	34.56	44.82	39.75	19.87	24.90	30.42	-
$\hat{\sigma}_X$	1.88	9.78	171.98	1.68	1.95	1.80	2.71	1.73	1.71	-
$\hat{\mu}_Y$	4.80	6.90	18.00	4.79	4.92	10.18	25.19	19.95	25.05	-
$\hat{\sigma}_Y$	1.88	9.78	171.98	1.68	1.95	1.80	2.71	1.73	1.71	-

Tabela 5.29: MCLUST: Klasyfikacja obserwacji

	numer skupienia								
nr obserwacji	1	2	3	4	5	6	7	8	9
1	0,076	0,915	0,009	0,000	0,000	0,000	0,000	0,000	0,000
2	0,942	0,057	0,002	0,000	0,000	0,000	0,000	0,000	0,000
3	0,971	0,027	0,002	0,000	0,000	0,000	0,000	0,000	0,000
4	0,959	0,036	0,006	0,000	0,000	0,000	0,000	0,000	0,000
5	0,562	0,400	0,037	0,000	0,000	0,000	0,000	0,000	0,000
6	0,971	0,027	0,002	0,000	0,000	0,000	0,000	0,000	0,000
7	0,538	0,449	0,012	0,000	0,000	0,000	0,000	0,000	0,000
8	0,621	0,361	0,017	0,000	0,000	0,000	0,000	0,000	0,000
9	0,320	0,672	0,008	0,000	0,000	0,000	0,000	0,000	0,000
10	0,952	0,032	0,016	0,000	0,000	0,000	0,000	0,000	0,000
...	...	...	...	...	...	...	...	...	...
31	0,000	0,979	0,021	0,000	0,000	0,000	0,000	0,000	0,000
32	0,000	0,988	0,012	0,000	0,000	0,000	0,000	0,000	0,000
33	0,000	0,991	0,009	0,000	0,000	0,000	0,000	0,000	0,000
34	0,000	0,957	0,043	0,000	0,000	0,000	0,000	0,000	0,000
35	0,000	0,980	0,020	0,000	0,000	0,000	0,000	0,000	0,000
36	0,000	0,991	0,009	0,000	0,000	0,000	0,000	0,000	0,000
37	0,000	0,986	0,014	0,000	0,000	0,000	0,000	0,000	0,000
38	0,000	0,973	0,027	0,000	0,000	0,000	0,000	0,000	0,000
39	0,000	0,964	0,036	0,000	0,000	0,000	0,000	0,000	0,000
40	0,000	0,977	0,023	0,000	0,000	0,000	0,000	0,000	0,000

Tabela 5.30: TwoStep: Klasyfikacja jednostek odstających

	klasyfikacja	
nr	rzeczywista	estymowana
144	5	-1
271	out	-1
273	out	-1
275	out	-1
276	out	-1
278	out	-1
279	out	-1
281	out	-1
282	out	-1
283	out	-1
284	out	-1
285	out	-1
286	out	-1
287	out	-1
288	out	-1
293	out	-1
294	out	-1
296	out	-1
297	out	-1
299	out	-1
300	out	-1
274	out	3
280	out	3
291	out	3
298	out	3
277	out	2
289	out	2
290	out	2
272	out	1
292	out	1
295	out	1

błędów związanych z identyfikacją outlierów, przedstawiamy podsumowanie popełnionych błędów ich klasyfikacji:

Tabela 5.31: TwoStep: Błędy decyzyjne przy identyfikacji jednostek odstających

		Stan faktyczny	
		$x \in out$	$x \notin out$
Decyzja	$x \in out$	20	1
	$x \notin out$	10	269

Miarą poprawności identyfikacji jest suma elementów na przekątnej unormowana przez ogólną liczebność populacji (=0.963). Warto zauważyć, że tylko jedna jednostka została niesłusznie uznana za odstającą, ale aż dokładniej jedna trzecia zbiorowości outlierów nie została poprawnie wykryta. Podkreślmy też, że uzyskany wynik jest spójny z identyfikacją dokonaną przez algorytm **Fanny** - por. Rozdział 2.

Rozdział 6

Aneks

6.1. Opis wybranych metod analizy skupień

6.1.1. Metody hierarchiczne

Metody hierarchiczne biorą swoją nazwę od hierarchicznej struktury, jaką tworzą skupienia. Struktura tej hierarchii wyznaczona jest przez schemat, który za chwilę opiszemy. Poniższy tok rozumowania oraz niektóre oznaczenia, które posłużą nam do opisu metod hierarchicznych pochodzą z artykułu Johnsona [37].

Załóżmy, że mamy n obserwacji, ciąg podziałów $C_0, C_1, \cdots, C_m$ zbioru oraz ciąg odpowiadających im wartości $\alpha_0, \alpha_1, \cdots, \alpha_m$. Będziemy je nazywać **współczynnikami fuzji** W przypadku gdy mamy do czynienia z metodami aglomeracyjnymi, ciągi te spełniają zależności:

$$C_{j-1} \leqslant C_j$$
$$\alpha_{j-1} \leqslant \alpha_j$$

W przypadku metod deglomeracyjnych należy po prostu odwrócić kierunek powyższych nierówności. Dla ustalenia uwagi, przyjmijmy, że będziemy zajmować sie teraz wyłącznie metodami aglomeracyjnymi.

Każdemu elementowi ciągu $C_0, C_1, \cdots, C_m$ przypisana jest **moc** jego grupowania, przy czym C_{j-1} jest słabsze od C_j. Te określenia wynikają z faktu, że kolejne grupowania tworzą *słaby* liniowy porządek. Inna interpretacja jest następująca intuicja oparta na definicji skupienia według Linga (zob. Rozdział 2). Klasy pierwszego poziomu są najsłabsze, ponieważ stopień podobieństwa $\alpha_0 = r = 0$ wyznacza długość trwania skupienia. Analogicznie, grupowania ostatniego poziomu są najmocniejsze, w tym sensie, że wartość r jest na tyle duża, że skupienie jest zachowane zawsze.

Taki schemat generuje specyficzny rodzaj **metryki**, zdefiniowanej jako $d(x, y) = \alpha_j$, gdzie j jest najmniejszym indeksem C_j ze zbioru (0,1,...,m) takim, że x, y należą do tego samego skupienia. Dowód, że jest to metryka można znaleźć w [37]. Zauważmy, że tak zdefiniowana odległość jest odpowiednikiem miary **strukturalnej równoważności** w dla modeli blokowych. Zasadnicza różnica między nimi a grupowaniem hierarchicznym polega na tym, że tutaj poziom podobieństwa jest zmienny w czasie przebiegu algorytmu.

W czasie przebiegu algorytmu następują kolejne przejścia od oryginalnej macierzy odległości do macierzy metryki zdefiniowanej powyżej. Algorytm działa następująco:

1. Zacznij od słabego grupowania C_0 oraz $\alpha_0 = 0$

2. Załóżmy, że mamy dany podział C_{j-1} i macierz odległości między obiektami (obserwacjami lub skupieniami). Niech teraz α_j będzie największym niezerowym elementem w tej macierzy (wartość zero oznacza identyczność obiektów lub przynależność do tego samego skupienia). Połącz te pary obiektów, dla których odległość jest równa α_j. W ten sposób tworzone jest skupienie C_j.

3. Aktualizuj macierze odległości dla kolejnych podziałów

4. Powtarzaj kroki (2) i (3) do momentu otrzymania najmocniejszego grupowania C_m.

Bezpośrednie konsekwencje opisanej powyżej metody są następujące:

1. Kolejne wartości współczynników fuzji α_j zwiększane są w kolejnych iteracjach, aż do momentu uzyskania skupienia składającego się z całego zbioru.

2. Im większy współczynnik, tym dane obiekty są do siebie podobne w mniejszym stopniu.

3. Jeśli dwa obiekty należały w kroku $k-1$ do danej klasy, to należą do niej również w kroku k tym.

Zdążyliśmy wcześniej zdefiniować odległość d między obserwacjami, nie wyjaśniając w jaki sposób określić odległość między obserwacją, a skupieniem.

W **metodzie najbliższego sąsiada** (ang. single linkage), odległość obserwacji $x$od skupienia C definiuje się jako odległość do najbliższej obserwacji z C. Formalnie:

$$d_{sing}(x, C) = \inf_{y \in C} d(x, y)$$

W **metodzie najdalszego sąsiada** (ang. complete linkage), odległość obserwacji $x$od skupienia C definiuje się jako odległość do najdalszego obserwacji z C. Formalnie:

$$d_{comp}(x, C) = \sup_{y \in C} d(x, y)$$

Metoda **średniego wiązania** (ang. average linkage) określa odległość jako średnią arytmetyczną odległości między x a obserwacjami z C

$$d_{ave}(x, C) = \frac{1}{n_C} \sum_{y_i \in C} d(x, y_i)$$

Metoda środka ciężkości (ang. centroid method) w pierwszej kolejności utożsamia każdy segment ze swoim środkiem ciężkości (zob. metoda K-średnich), co pozwala określić odległość między klasami, jako odległość między ich centroidami:

$$d_{cen}(x, C) = d(x, \mu_C)$$

Warto jeszcze omówić krótko **metodę Warda**, która różni się koncepcyjnie od omówionych wcześniej. Zamiast macierzy odległości wykorzystuje ona kryterium sumy kwadratów odległości wewnątrzgrupowych, czyli WSS. W pierwszym kroku sumaryczna wartość kryterium jest równa zero, ponieważ każda obserwacja tworzy oddzielne skupienie, co implikuje zerowe zróżnicowanie zmiennych wewnątrz grup. W następnym kroku porównywane są

wszystkie $\binom{n}{2}$ dwuelementowe podzbiory. Skupienie tworzą te, które zapewniają minimalną ogólną wartość sumy kwadratów wewnątrzgrupowych. W kolejnych krokach aktualizowane są informacje na temat wektora średnich w każdym skupieniu i łączone są te skupienia, które minimalizują wspomniane kryterium.

6.1.2. Algorytm K-średnich

Algorytm K-średnich (ang. K-means) należy do szerokiej klasy algorytmów optymalizacyjnych. Wymaga określenia przez użytkownika odpowiedniej liczby skupień, w ramach których minimalizowane jest odpowiednie kryterium. Według kryteriów zdefiniowanych w Rozdziale 2 należy on do metod niehierarchicznych, co oznacza, że nie tworzy żadnej struktury skupień tj. żadne ze skupień nie jest zawarte w innym. K-średnich nie jest zatem ani deglomeracyjny ani aglomeracyjny. Uzyskany podział jest ścisły w sensie klasycznym tj. każda obserwacja należy do jednego i tylko jednego skupienia.

Ogólna postać szerokiej klasy algorytmów K-means wygląda następująco:

1. Zdefiniuj funkcję celu i warunki ograniczające na liczbę klas
2. Ustal poziom tolerancji ϵ dla zmiany wartości funkcji celu
3. Wyznacz rozwiązanie początkowe
4. Dla każdego obiektu wyznacz najbliższe skupienie
5. Zmień przyporządkowanie obserwacji, tak aby poprawić wartość funkcji celu
6. Zakończ procedurę, gdy zmiana wartości funkcji celu jest nie większa od ϵ

Tutaj ograniczymy się jedynie do najbardziej powszechnej metody. Jako funkcję celu przyjmuje się sumę kwadratów odchyleń wewnątrzgrupowych WSS. Zakres tolerancji zazwyczaj wbudowany jest w oprogramowanie, przy czym jego dokładność ustalana jest na poziomie 10^{-5}, co zapewnia dość dobrą dokładność. Wyznaczenie warunków początkowych polega na określeniu k startowych centroidów (lub równoważnie środków ciężkości lub wektorów średnich warunkowych w grupach), co wyjaśnia etymologię nazwy algorytmu. Wybór początkowych centroidów opiera się o pewien schemat (np. branych jest pierwszych k jednostek) lub ma charakter losowy (wybieranych jest k liczb spośród n obserwacji zgodnie z rozkładem jednostajnym). W niektórych pakietach (np. SPSS) użytkownik może sam zdefiniować współrzędne punktów startowych lub wczytać je z poprzednich analiz.

Kroki (4) i (5) stanowią z kolei podstawę działania algorytmu i wykonywane są naprzemiennie aż do uzyskania efektu opisanego przez krok (6).

Po wczytaniu wyjściowych współrzędnych centroidów, dla każdej pary złożonej z obiektu i centroidu obliczane są odległości euklidesowe. Na ich podstawie dany obiekt przyłączany jest do najbliższego mu centroidu. W ten sposób k centroidów dzieli przestrzeń na k segmentów. W ostatniej fazie aktualizowane są współrzędne wektora średnich dla każdego skupienia. Aktualizacja może dokonywać się już po zaklasyfikowaniu pojedynczej jednostki (tzw. ruchome średnie - ang. running means) lub po wyczerpaniu wszystkich obserwacji (wersja klasyczna). Algorytm kończy swój przebieg w momencie, gdy liczba iteracji przekroczy ustaloną wartość lub zmiana w centrach skupień będzie mniejsza od wyznaczonego poziomu ϵ.

W niektórych pakietach (np. SPSS) algorytm k-średnich znany jest jako szybka klasyfikacja (ang. quick-cluster) przede wszystkim ze względu na dość niską złożoność obliczeniową, co jest ogromną zaletą dla bardzo dużych zbiorów danych. W przeciwieństwie do algorytmów

hierarchicznych (które za chwilę dokładniej przedstawimy) metoda K-średnich nie porównuje parami wszystkich obiektów lecz ogranicza się do pewnego zakresu podzbiorów, przez co działa znacznie szybciej niż algorytmy hierarhiczne.

Szybkość przebiegu jest ceną, jaką trzeba zapłacić za brak gwarancji globalnej optymalności rozwiązania. Na lokalną optymalność rozwiązania ma wpływ wybór punktów startowych. Algorytm zwraca różne wyniki dla różnych permutacji wierszy surowej macierzy danych. Algorytm jest podatny na obecność jednostek odstających. Gdy ich odsetek przekracza pewną wartość, pojawia się ryzyko zaburzonego wyniku, ponieważ taka jednostka ma duże prawdopodobieństwa trafienia do zbioru punktów startowych. Numeryczne eksperymenty [58] pokazują, że nawet dla niewielkich zbiorów ($n = 200, k = 8$) liczba różnych rozwiązań sięga kilku tysięcy. Dlatego wielu statystyków próbowało rozwinąć inteligentne metody wyznaczenia punktów startowych. Wśród 12 zebranych propozycji znajdowały się m.in. metody zaimplementowane w programach SPSS 12.0 i SAS/STAT 9.1, jak również "chałupnicza" metoda autorstwa Steinleya polegająca na wielokrotnym powtarzaniu algorytmu dla różnych punktów startowych. Wyniki okazały się zaskakujące. Najgorzej poradziły sobie algorytmy wbudowane w SPSS i SAS, a najlepiej metoda zaproponowana przez Stenleya!

Kolejnym słabym punktem metody K-średnich jest to, co stanowi jej założenie tzn. konieczność wyboru liczby klas przez użytkownika. Gdy nie dysponujemy odpowiednią teorią lub względami praktycznymi, problem liczby skupień sprowadza się do analogicznego problemu co wybór punktów startowych. Również tutaj sensownym rozwiązaniem jest wygenerowanie wyników dla pewnego przedziału dopuszczalnej liczby skupień i wybrania optymalnego. O wybranych kryteriach optymalności możemy przeczytać w Rozdziale 2.

6.1.3. Metoda grupowania dwustopniowego (ang. Two-Step Cluster)

Twórcy SPSS zaznaczają [57], że klasyczne metody analizy skupień, do których zalicza się również wyżej opisane, działają efektywnie na *małych* zbiorach obserwacji, jednak nie radzą sobie z *dużymi* bazami danych. Autorzy nie precyzują dokładnie, co oznaczają konkretnie podane określenia wielkości zbioru. Jeśli im ufać, grupowanie dwustopniowe (ang. Two-Step clustering) łączy w sobie zalety tradycyjnych metod, jednocześnie nie powielając wad swoich poprzedników. Nazwa algorytmu wywodzi się z faktu, że wyróżnia się w nim dwie fazy działania.

Krok 1 - Faza pre-klasyfikacyjna

Pierwszy krok polega na budowaniu z pojedynczych obserwacji struktury hierarhicznej w postaci drzewa, które nosi nazwę **Cluster-Feature Tree**. Stosując terminologię informatyczną, drzewo jest strukturą składającą się z **wierzchołków** oraz **gałęzi**. Drzewo rozrasta się w odwrotnym kierunku niż podpowiada nam intuicja. W ten sposób szczytowym wierzchołkiem jest **korzeń**, a najniższe wierzchołki to **liście**. Do korzenia nie dochodzi żadna gałąź, natomiast z liści żadna gałąź nie wychodzi.

Na początku, drzewo składa się tylko z jednego elementu, który jest jednocześnie korzeniem i liściem. Podobnie jak w metodach aglomeracyjnych taka pojedyncza obserwacja traktowana jest jako oddzielne skupienie. Każda kolejna obserwacja, która jest dołączana do drzewa porównywana jest z poprzednimi, zaczynając od korzenia, poprzez wierzchołki pośrednie, aż zostanie zaklasyfikowana do odpowiedniego liścia lub utworzy ona nowy liść. Każdy liść reprezentuje oddzielne skupienie. Zawiera informację na temat liczby obserwacji oraz parametrów, którego opisują (np. średnia, wariancja lub częstość kategorii w przypadku zmiennych dyskretnych).

Porównywanie obiektów odbywa się w oparciu o metrykę euklidesową w przypadku zmiennych ciągłych lub **metrykę największej wiarogodności** w przypadku zmiennych dyskretnych. Pomysł na wykorzystanie tej ostatniej bierze się z metod konstrukcji **drzew filogenetycznych** wykorzystywanych przede wszystkich w genetyce (por.?). Podobnie jak w metodzie Warda przy grupowaniu hierarchicznym dla każdej pary obiektów, które mają utworzyć jedno skupienia obliczane jest kryterium, którego wartość należy zminimalizować. Jak podaje Liu, tym kryterium jest minimalizacja wyrażenia:

$$d(A,B) = \log L(A) + \log L(B) - \log(A,B) \tag{6.1}$$

Gdzie np. $L(A)$ oznacza wartość funkcji wiarogodności dla skupienia A. Natomiast $L(A,B)$ oznacza wartość funkcji wiarogodności dla połączonych skupień A i B.

Pierwsza faza algorytmu kończy się w momencie pogrupowania wszystkich obserwacji. W ten sposób dokonuje się redukcja zbioru obserwacji do *mini*-skupień, które trakowane są jako obserwacje drugiego poziomu i biorą udział w analizie skupień w kolejnym kroku.

Krok 2 - Faza grupowania skupień

Drugi etap zaczyna się od stworzenia nowego roboczego zbioru danych, którego obserwacjami są liście uzyskane w fazie pierwszej, a zmiennymi charakterystyki tychże liści. Następnie przeprowadzana jest analiza skupień uzyskanych za pomocą jednej z klasycznych metod. Najczęściej wykorzystuje się procedury hierarchiczne jako nie wymagające narzucenia konkretnej liczby klas [57]. Wybór odpowiedniej liczby segmentów dokonywany jest automatycznie w oparciu o kryterium informacyjne BIC.

Do podstawowych zalet algorytmu zalicza się możliwość użycia zmiennych z różnego poziomu pomiaru (zarówno ciągłego, jak i dyskretnego). Jest to możliwe dzięki zastosowaniu uniwersalnej metryki największej wiarogodności. Inną zaletą jest automatyczny wybór liczby skupień bez konieczności ingerencji użytkownika. Dodatkowo, algorytm ma wbudowaną obsługę identyfikacji jednostek odstających z możliwością ustalenia procentowego pułapu ich obecności w próbie.

O ostatecznym kształcie zbioru decyduje krok drugi czyli grupowanie skupień. Bez względu na dokładną budowę drzewa w pierwszym kroku, misterna konstrukcja może ulec zmianie w drugiej fazie. Innymi słowy, algorytm posiada te same przypadłości, co metody hierarchiczne.

Po zakończeniu pierwszej fazy użytkownik nie może zweryfikować jej rezulatów i podjąć decyzji odnośnie konieczności przeprowadzania drugiej fazy. Nie ma też wpływu na wybór algorytmu hierarchicznego w drugim etapie.

Ponadto, Two-Step miał być pierwszym algorytmem SPSS wykorzystującym metody probabilistyczne. W praktyce ich obecność ogranicza się do zastosowania metody największej wiarogodności do zdefiniowania metryki dla zmiennych dyskretnych oraz obliczenia BIC. Jednak sposób przyporządkowania obserwacji do skupień nadal pozostaje deterministyczny tj. każda jednostka należy do dokładnie jednego skupienia z prawdopodobieństwem równym 1. Co więcej, metoda nie oferuje żadnej informacji na temat błędu klasyfikacji ani stopnia dopasowania modelu do danych. Trudno też w ogóle mówić o modelu jako o całości, gdyż elementy statystyki pojawiają się jedynie w pierwszej fazie algorytmu.

6.2. Opis zbiorów ilustracyjnych

Czysto teoretyczne omówienie pojawiających się w niniejszej części pracy różnego rodzaju współczynników, indeksów, miar dopasowania, kryteriów itp. związanych z analizą skupień byłoby o wiele mniej interesujące, gdyby pozbawione było ilustracji na konkretnych przykładach. Dlatego postanowiliśmy wygenerować dwa proste, dwuwymiarowe zbiory danych wyraźnie różniące się pod względem struktury.

Pierwszy z nich nazwaliśmy **amorficznym** (oznaczać go będziemy literą $\mathcal{A}$) ponieważ nie da się w nim wydzielić wyraźnie rozłącznych skupień. Drugi zbiór nazwaliśmy **skrystalizowanym** i oznaczać go będziemy literą $\mathcal{S}$ ponieważ można w nim wyróżnić trzy lub dziewięć rozłącznych skupień, w zależności od stopnia dokładności zastosowanej metody.

Obydwa zbiory zostały wygenerowane za pomocą **pakietu R**. Zbiór A liczy 300 obserwacji pochodzących z jednego dwuwymiarowego standardowego rozkładu normalnego. Zbiór S również składa się z 300 obserwacji, z czego 270 pochodzi z 9 różnych klas, którym odpowiadają różne parametry rozkładu dwuwymiarowego rozkładu normalnego. Parametry te można odczytać z postaci kodu źródłowego, który bazuje na następującym układzie równań [**?**, s.68]:

$$X = aU + bV + h$$
$$Y = cU + dV + k$$

Jeśli $U, V \sim N(0,1)$ wówczas wektor (X, Y) ma rozkład normalny o wartości oczekiwanej (h, k) i macierzy kowariancji

$$\begin{bmatrix} a^2 + b^2 & ac + bd \\ ac + bd & a^2 + b^2 \end{bmatrix}$$

Jeśli $ac + bd = 0$ wówczas zmienne wchodzące w skład wektora są niezależne.

W zbiorze S dodatkowo znajduje 30 obserwacji pochodzących z dwuwymiarowego rozkładu jednostajnego o nośniku $(0, 50)$ dla zmiennej X oraz $0, 30$ dla zmiennej Y. Jest to tzw. **szum** lub zbiór jednostek odstających, który ma za zadanie utrudniać pracę algorytmu przy określaniu skupień.

6.3. Kod źródłowy R do generowania zbiorów ilustracyjnych

Poniżej znajduje się kod źródłowy generujący rozkłady pojawiające się pracy. Należy jednak pamiętać, że losowość w generowaniu zbioru może powodować pewne rozbieżności prezentowanych przez nas wyników z uzyskanymi przez Czytelnika. W jakim stopniu te różnice powinny występować jest ciekawym problemem statystycznym i został on poruszony w pracy, w szczególności w rozdziale poświęconemu analizie skupień za pomocą modeli probabilistycznych. Dlatego aby uzyskać zgodność i możliwość porównania w wyników w Aneksie można skontaktować się z autorem tej pracy: pzimol@gmail.com

```
#załadowanie biblioteki

library(MASS)

#Wygenerowanie dwóch zbiorów: A - amorficzny, S - skrystalizowany
```

```
A=cbind(rnorm(300,0,1)+rnorm(300,0,1), rnorm(300,0,1)-rnorm(300,0,1))

memb=rep(1:10, each=30)

c1 = cbind(rnorm(30,0,1)+rnorm(30,0,1) +5, rnorm(30,0,1)-rnorm(30,0,1) +5)
c2 = cbind(rnorm(30,0,1)+rnorm(30,0,1) +15, rnorm(30,0,1)-rnorm(30,0,1) +5)
c3 = cbind(rnorm(30,0,1)+rnorm(30,0,1) +10, rnorm(30,0,1)-rnorm(30,0,1) +10)
c4 = cbind(rnorm(30,0,1)+rnorm(30,0,1) +35, rnorm(30,0,1)-rnorm(30,0,1) +5)
c5 = cbind(rnorm(30,0,1)+rnorm(30,0,1) +45, rnorm(30,0,1)-rnorm(30,0,1) +5)
c6 = cbind(rnorm(30,0,1)+rnorm(30,0,1) +40, rnorm(30,0,1)-rnorm(30,0,1) +10)
c7 = cbind(rnorm(30,0,1)+rnorm(30,0,1) +20, rnorm(30,0,1)-rnorm(30,0,1) +25)
c8 = cbind(rnorm(30,0,1)+rnorm(30,0,1) +25, rnorm(30,0,1)-rnorm(30,0,1) +20)
c9 = cbind(rnorm(30,0,1)+rnorm(30,0,1) +30, rnorm(30,0,1)-rnorm(30,0,1) +25)
c10= cbind(runif(30,0,50), runif(30,0,30))

Sraw=rbind(c1,c2,c3,c4,c5,c6,c7,c8,c9,c10)
S1=cbind(Sraw,memb)
classif=S1[,3]
S=S1[,-3]

#Wyświetl macierze danych surowych

A
S

#Przedstaw wykresy rozrzutu

par(mfrow=c(1,2))
plot(A)
plot(S)

#scaterplot
```

Spis literatury

[1] A. Agresti. *Statistical Methods for the Social Sciences*. Pearson Prentice Hall, 2002.

[2] M. Aitkin, D. Anderson, and J. Hinde. Statistical modeling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, 144(41):419–461, 1981.

[3] M. Aitkin and D. Rubin. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(1):67–75, 1985.

[4] P. Arabie, W. DeSarbo D. Carroll, and J. Wind. Overlapping clustering: A method of product positioning. *Journal of Marketing Research*, 18:310–317, 1981.

[5] G. Arminger and U. Kusters. Construction principles for latent trait models. *Sociological Methodology*, 19:369–393, 1989.

[6] K. Bailey. Cluster analysis. *Sociological Methodology*, 61:59–128, 1975.

[7] F. Baker and L. Hubert. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38, 1975.

[8] H. Banaszak. Skalowanie - hasło w encyklopedii socjologicznej pwn.

[9] H. Banaszak. Identyfikacja struktury jako problem decyzyjny. *Prakseologia*, 95-96(3-4):1–19, 1985.

[10] J.D. Banfield and A. Raftery. Model based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

[11] V. Barnett. The study of outliers: Purpose and model. *Applied Statistics*, 27(3):242–250, 1978.

[12] D.J. Bartholomew. Scaling unobservable constructs in social science. *Applied Statistics*, 47(1):1–13, 1998.

[13] K.E. Basford and G.T. McLachlan. Likelihood estimation with normal mixtures models. *Applied Statistics*, 34(3):282–289, 1985.

[14] P. Bekker and T. Wansbek. *A companion to theoretical econometrics*, chapter 7. Identification in Parametric Models. Wiley-Blackwell, 2003. dostępna na http://www.blackwellreference.com.

[15] H. Bock. On some significance tests in cluster analysis. *Journal of Classification*, 2:77–108, 1985.

[16] C. Bruni and G. Koch. Identifiability of continous mixtures of unknown gaussian distributions. *The Annals of Probability*, 13(4):1341–1357, November 1985.

[17] R.S. Burt. Models of network structure. *Annual Review of Sociology*, 6:79–141, 1980.

[18] G.A. Churchill. *Badania marketingowe. Podstawy metodologiczne.* PWN, Warszawa, 2002.

[19] R.M. Cormack. A review of classification. *Journal of the Royal Statistical Society Series A (General)*, 134(3):321–367, 1971.

[20] A. Dasgupta and A. Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302, March 1998.

[21] L. Davies and U. Gather. Identification of multiple outliers. *Journal of the American Statistical Association*, 88(423):782–792, 1993.

[22] F. Dellaert. The expectation maximization algorithm. Technical Report GIT-GVU-02-20, College of Computing, Georgia Institute of Technology, February 2002.

[23] A.P. Dempster, N.M. Laird, and Rubin D.B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[24] B. Efron. Bootstrap methods: Another look at the jacknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[25] B.S. Everitt. Unresolved problems in cluster analysis. *Biometrics*, 35(1):169–181, 1979.

[26] B.S. Everitt. A monte carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, 16:171–180, 1981.

[27] B.S. Everitt, S. Landau, and M. Leese. *Cluster Analysis.* Arnold. The Oxford University Press, London, fourth edition, 2001.

[28] C. Fraley and A. Raftery. How many clusters? which clustering method? answer via model based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[29] C. Fraley and A. Raftery. *The MCLUST Package for R. Model-Based Clustering. Normal Mixture Modeling.* http://www.stat.washington.edu/fraley/mclust, 2.1-10.3 edition, March 2009.

[30] M.T. Gallegos and G. Ritter. Robust method for cluster analysis. *The Computer Journal, The Annals of Statistics*, 3(1):347–380, 2005.

[31] W.A. Gibson. Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika*, 24(3):229–251, 1959.

[32] L. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231, 1974.

[33] A.D. Gordon. A review of hierarchical classification. *Journal of the Royal Statistical Society. Series A (General)*, 150(2):119–137, 1987.

[34] P. Green and V. Rao. Note on proximity measures and cluster analysis. *Journal of Marketing Research*, 6(3):359–364, 1969.

[35] W. Guzicki and P. Zakrzewski. *Wykłady ze wstępu do matematyki. Wprowadzenie do teorii mnogości.* PWN, Warszawa, 2005.

[36] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, New York, 2001.

[37] U. Hoehle. On the fundamentals of fuzzy sets theory. *Journal of Mathematical Analysis and Applications*, 201(0285):786–826, 1996.

[38] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[39] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering. a review. *ACM Computing Surveys*, 31(3):265–323, 1999.

[40] H. Jeffreys. *The Theory of Probability.* Oxford University Press, 3e edition, 1961.

[41] S.C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[42] R. Kaufman. Issues in multivariate cluster analysis. *Sociological Methods and Research*, 13(4):467–486, 1985.

[43] T.D. Klastorin. Assessing cluster analysis results. *Journal of Marketing Research*, 20(1):92–98, 1983.

[44] J. Komorek. Jednowymiarowe metody skalowania jednowymiarowego. Master's thesis, Uniwersytet Warszawski. Instytut Socjologii, Warszawa, czerwiec 2006. praca napisana pod kierunkiem dra. H.Banaszaka.

[45] J. Koronacki and J. Ćwik. *Statystyczne systemy uczące się.* EXIT. Akademicka Oficyna Wydawnicza, Warszawa, 2008.

[46] M. Krzyśko, W. Wołyński, T. Górecki, and M. Skorzybut. *Systemy uczące się Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości.* Wydawnictwa Naukowo-Techniczne, Warszawa, 2008.

[47] P.F. Lazarsfeld. Recent development in latent structure analysis. *Sociometry and the Science of Man*, 18(4):391–403, 1955.

[48] P.F. Lazarsfeld. *Latent Structure Analysis and Test Theory*, chapter 2. The M.I.T. Press, 1968. w: Readings in Mathematical Social Science.

[49] F.R. Ling. A probability theory of cluster analysis. *Journal of the American Statistical Association*, 68, 1973.

[50] Y. Lo, N.R. Mendell, and D. Rubin. Testing the number of components in a normal mixture. *Biometrika*, 88(3):767–778, September 2001.

[51] T. Marek. *Analiza skupień w badaniach empirycznych. Metody SAHN.* PWN, Warszawa, 1989.

[52] G.J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36(3):318–324, 1987.

[53] G. Milligan. Monte carlo study of criterion measures for cluster analysis. *Psychometrika*, 46(2):187–199, 1981.

[54] G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in data set. *Psychometrika*, 50(2):159–179, 1985.

[55] W. Navidi. Graphical illustration of the em algorithm. *The American Statistician*, 51(1):29–31, 1997.

[56] R. Peck, L. Fisher, and J. Ness. Approximate confidence intervals for the number of clusters. *Journal of the American Statistical Association*, 84(405):184–191, 1989.

[57] G. Punj and D. Steward. Cluster analysis in marketing research review and suggestions for application. *Journal of Marketing Research*, 20(2):134–148, 1983.

[58] R.A. Redner and H.F. Walker. Mixture densities maximum likelihood and the em algorithm. *Society for Industrial and Applied Mathematics*, 26(2):195–239, 1984.

[59] W.W. Sawyer. *Droga do matematyki współczesnej*. Omega. Wiedza Powszechna, Warszaw, pierwsze edition, 1974.

[60] S.D. Silvey. *Wnioskowanie statystyczne*. PWN, 1978.

[61] A. Smith and D. Spiegelhalter. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society. Section B*, 42:213–220, 1980.

[62] P.H. Sneath. Some statistical problems in numerical taxonomy. *The Statistician*, 17(1):1–12, 1967.

[63] SPSS. The spss twostep cluster component. White paper, The SPSS Inc., http://www.spss.com, 2001.

[64] D. Steinley and M. Brusco. Initializing k-means batch clustering. a critical evaluation of several techniques. *Journal of Classification*, 24:99–118, 2007.

[65] C. Sugar and G. James. Finding the number of clusters in a data set an information theoretic approach. *Journal of the American Statistical Association*, 98(463):750–763, 2003.

[66] H. Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248, March 1961.

[67] H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, December 1963.

[68] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Methodology)*, 63(2):411–423, 2001.

[69] O. Wagner. Analiza klas ukrytych p.f.lazarsfelda. wybrane zagadnienia. Master's thesis, Uniwersytet Warszawski. Instytut Socjologii, Warszawa, 2004. praca napisana pod kierunkiem prof. dr hab. Grzegorza Lissowskiego.

[70] J. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 55:329–350, 1970.

[71] M. Xiao and D. Dyk. The em algorithm. an old folk song sung to a fast tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):511–567, 1997.

[72] K. Yeung and W. Ruzzo. Details of the adjusted rand index and clustering algorithms, supplement to the paper.

[73] J. Yu, J. Amores, S. Nicu, and Q. Tian. A new study on distance metrics as similiraty measurement. In *IEEE International Conference on Multimedia and Expo*, pages 533–536, http://dare.uva.nl/record/221614, 2006. University of Amsterdam.

[74] R. Zieliński. *Siedem wykładów wprowadzających do statystyki matematycznej.* Instytut Matematyki PAN, Warszawa, 2004. dostępny na http://www.impan.gov.pl/ rziel/7ALL.pdf.