Zeszyty Naukowe Wydziału Informatyki
Wyższej Szkoły Informatyki Stosowanej i Zarządzania
„Informatyka Stosowana”
Nr 1/2014

KOMPUTEROWE MODELOWANIE WIELOWYMIAROWYCH DANYCH Z WYKORZYSTANIEM KOPUŁ

Hubert Czobodziński

Wyższa Szkoła Informatyki Stosowanej i Zarządzania
ul. Newelska 6, 01–447 Warsaw, Poland

Streszczenie

Przedmiotem pracy jest implementacja metody modelowania cenzorowanych danych wielowymiarowych z wykorzystaniem kopuły przedstawionej w artykule Chen i Fan (2007). Praca zawiera podstawowe informacje o kopułach, zagadnieniach analizy przeżycia i niezawodności oraz omówienie wybranych aspektów modelowania danych wielowymiarowych. Przestawiono w niej szczegóły implementacji jednej z metod modelowania danych cenzorowanych oraz wyniki doświadczeń numerycznych, przeprowadzonych przy użyciu sztucznie wygenerowanych danych losowych.

Słowa kluczowe: kopuły, dane wielowymiarowe, dane cenzorowane, modelowanie, niezawodność

1 Wstęp

Wiele dziedzin, w których wykorzystywane są narzędzia statystyczne boryka się z problemem modelowania danych empirycznych pochodzących z rozkładów daleko odbiegających od dwuwymiarowego rozkładu normalnego.

Problemy o takim charakterze pojawiają się między innymi w geostatystyce, finansach, ubezpieczeniach czy badaniach epidemiologicznych (Frees, Valdez, 1998). Ostatni z przykładów jest szczególnie istotny – także z uwagi na tematykę niniejszej pracy – gdyż pojawiają się w nim dodatkowe kwestie, charakterystyczne dla analizy przeżycia i niezawodności.

Jedną z nadziei na poprawę tego stanu są *kopuły* – narzędzie do modelowania rozkładów wielowymiarowych o skomplikowanych strukturach zależności, zdobywające coraz większą popularność i z powodzeniem stosowane w wymienionych wyżej dziedzinach.

2 Kopuły

2.1 Definicja

Definicja 1 *Dwuwymiarowa kopuła to funkcja* $C : [0,1]^2 \to [0,1]$ *o następujących własnościach:*

1. *Dla dowolnych* $u, v \in [0,1]$

$$C(u,0) = C(0,v) = 0 \tag{1}$$

oraz

$$C(u,1) = u \ \wedge \ C(1,v) = v. \tag{2}$$

2. *Dla dowolnych* $u_1, u_2, v_1, v_2 \in [0,1]$*, takich że* $u_1 \leq u_2$ *i* $v_1 \leq v_2$

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0. \tag{3}$$

Wprost z definicji wynika, że kopuła jest dystrybuantą łączną dwuwymiarowego rozkładu o jednostajnych rozkładach brzegowych, określonego na kwadracie jednostkowym.

Kolejnym wnioskiem z definicji jest następująca nierówność, prawdziwa dla każdej kopuły C i każdego $(u,v) \in [0,1]^2$:

$$W(u,v) \leq C(u,v) \leq M(u,v). \tag{4}$$

gdzie $W(u,v) = \max(u+v-1, 0)$ oraz $M(u,v) = min(u,v)$. Funkcje W i M (również będące kopułami) nazywane są granicami Frécheta-Hoeffdinga.

2.2 Twierdzenie Sklara

Jednym z najistotniejszych elementów teorii kopuł i fundamentem dużej części przypadków zastosowania jej w statystyce, jest twierdzenie Sklara.

Twierdzenie 1 (Sklar) *Niech* $F(x,y)$ *będzie dystrybuantą łączną rozkładu dwuwymiarowego o dystrybuantach brzegowych* $F_x(x)$ *i* $F_y(y)$*. Istnieje kopuła* C*, taka że:*

$$F(x,y) = C(F_x(x), F_y(y)). \tag{5}$$

Jeśli F_x *i* F_y *są ciągłe to* C *jest jednoznacznie określona. W przeciwnym przypadku* C *jest jednoznacznie określona na iloczynie kartezjańskim przeciwdziedzin* F_x *i* F_y*. Prawdziwe jest również twierdzenie odwrotne: jeśli* C *jest kopułą a* F_x *i* F_y *dystrybuantami to funkcja* F *zdefiniowana przez (5) jest dystrybuantą łączną rozkładu dwuwymiarowego o dystrybuantach brzegowych* F_x *i* F_y*.*

Twierdzenie to pojawiło się po raz pierwszy w pracy Sklara (1959). Nazwa *kopuła*[1] ma obrazować sposób w jaki funkcja ta łączy dystrybuantę rozkładu łącznego z rozkładami brzegowymi.

Twierdzenie Sklara, z punktu widzenia statystyki, czyni kopuły narzędziem pozwalającym na badanie struktury zależności wielowymiarowych zmiennych losowych w oderwaniu od ich rozkładów brzegowych.

2.3 Kopuły archimedesowskie

Nazwa kopuły archimedesowskie określa kopuły postaci:

$$C(u,v) = \phi^{[-1]}\left(\phi(u) + \phi(v)\right), \tag{6}$$

gdzie $\phi : [0,1] \to [0,\infty)$ jest ciągłą, wypukłą funkcją malejącą, spełniającą warunek $\phi(1) = 0$, a $\phi^{[-1]}$ jej pseudo-odwrotnością:

$$\phi^{[-1]} = \begin{cases} \phi^{-1}(t) & 0 \le t \le \phi(0) \\ 0 & \phi(0) < t \end{cases}. \tag{7}$$

Funkcja ϕ nazywana jest generatorem kopuły.

Obszerny przegląd kopuł archimedesowskich wraz z wykresami losowych próbek danych znaleźć można w pracy Armstrong (2003).

Kopuły archimedesowskie są szczególnie interesujące z punktu widzenia zastosowań praktycznych, ponieważ przy prostej formie obejmują wiele różnych struktur zależności zmiennych.

2.4 Dystrybuanta kopuły

Dwuwymiarowa kopuła może być traktowana jako zmienna losowa:

$$C : [0,1]^2 \to [0,1]. \tag{8}$$

Funkcja:

$$K(v) = P(C(X,Y) \le V) \tag{9}$$

jest dystrybuantą zmiennej losowej C. Dystrybuanta ta, w przypadku kopuł archimedesowskich może być wyrażona za pomocą generatora kopuły:

$$K(v) = v - \frac{\phi(v)}{\phi'(v)}. \tag{10}$$

Ponadto, jak wskazują autorzy pracy Genest, Rivest (1993), w tej klasie kopuł istnieje wzajemnie jednoznaczne odwzorowanie między kopułą a jej dystrybuantą, co umożliwia – w praktycznych zastosowaniach – wykorzystanie $K(v)$ do estymacji parametrów kopuł archimedesowskich.

[1] łaciński rzeczownik *copula* oznacza łączenie lub wiązanie

2.5 Tau Kendalla

Współczynnik τ Kendalla jest jedną z miar monotonicznej zależności między zmiennymi, zdefiniowaną jako (Kendall, 1948):

$$\tau = P\left[(x_1 - x_2)(y_1 - y_2) > 0\right] - P\left[(x_1 - x_2)(y_1 - y_2) < 0\right]. \tag{11}$$

Warto zwrócić uwagę, że o ile powszechnie stosowany współczynnik korelacji liniowej określa jedynie siłę liniowej zależności między zmiennymi, o tyle τ Kendalla pozwala na badanie także bardziej złożonych form zależności monotonicznej.

Jeśli X i Y są zmiennymi losowymi o charakterze ciągłym, to zależność (11) może być wyrażona w terminach kopuł:

$$\tau = 4 \int_0^1 \int_0^1 C_{X,Y}(u, v) dC_{X,Y}(u, v) - 1, \tag{12}$$

co w przypadku kopuł archimedesowskich pozwala na wyznaczenie τ Kendalla na podstawie generatora kopuły (Genest, MacKay, 1986):

$$\tau = 4 \int_0^1 \frac{\phi(v)}{\phi'(v)} dv + 1. \tag{13}$$

3 Analiza przeżycia i niezawodności

Analiza przeżycia i niezawodności jest szczególnym działem statystyki. Szczególnym ze względu na zagadnienia, którymi się zajmuje, charakterystyczne problemy w nich występujące oraz specyficzne narzędzia.

Obszar zainteresowań analizy przeżycia i niezawodności to zmienne losowe odpo-wiadające czasowi upływającemu do momentu wystąpienia jakiegoś zdarzenia. Zda-rzeniem takim może być śmierć osobnika w wybranej populacji (stąd nazwa – analiza przeżycia), ale też moment wystąpienia awarii pewnego urządzenia, czy czas pozostawania bezrobotnym po utracie pracy.

3.1 Funkcja przeżycia

Podstawowym obiektem zainteresowania analizy przeżycia i niezawodności jest *funk-cja przeżycia*:

$$\bar{F}(x) = P(X > x). \tag{14}$$

Jest ona ściśle związana z dystrybuantą zmiennej losowej:

$$\bar{F}(x) = 1 - F(x) \tag{15}$$

i określa prawdopodobieństwo wystąpienia zdarzenia po czasie dłuższym niż x (np. prawdopodobieństwo przeżycia czasu dłuższego niż x).

Dla dwuwymiarowej zmiennej losowej funkcja przeżycia jest określona jako

$$\bar{F}(x,y) = P(X > x, Y > y), \tag{16}$$

a jej związek z dystrybuantą jest następujący

$$\bar{F}(x,y) = 1 - F_x(x) - F_y(y) + F(x,y), \tag{17}$$

gdzie $F_x(x)$ i $F_y(y)$ są dystrybuantami brzegowymi.

Granice dwuwymiarowej funkcji przeżycia:

$$\bar{F}_x(x) = \lim_{y \to -\infty} \bar{F}(x,y) \tag{18}$$

oraz

$$\bar{F}_y(y) = \lim_{x \to -\infty} \bar{F}(x,y) \tag{19}$$

są jednowymiarowymi funkcjami przeżycia.

3.2 Funkcja hazardu

Specyfika niektórych zastosowań analizy przeżycia i niezawodności sprawia, że ważne staje się "negatywne" uzupełnienie informacji niesionej przez funkcję przeżycia. Realizowane jest ono przez funkcję hazardu:

$$\lambda(x)dx = P(x \leq X < x + dx | T \leq x) = -\frac{\bar{F}'(x)}{\bar{F}(x)}dx. \tag{20}$$

Funkcja hazardu określa prawdopodobieństwo zajścia zdarzenia w jednostce cza-su, pod warunkiem, że w danym przypadku do zdarzenia jeszcze nie doszło. Alternatywną reprezentacją funkcji hazardu jest skumulowana funkcja hazardu:

$$\Lambda(x) = \int_0^x h(t)dt = -\log(\bar{F}(x)). \tag{21}$$

3.3 Obserwacje ucięte

Cechą charakterystyczną zjawisk, którymi zajmuje się analiza przeżycia i niezawodności jest występowanie obserwacji uciętych (cenzorowanych). Pojawiają się one – w najogólniejszym przypadku – wtedy, gdy wiadomo jedynie, że prawdziwy czas do wystąpienia pewnego zdarzenia jest nie mniejszy niż czas zaobserwowany. Najczęstszą przyczyną cenzorowania jest konieczność zakończenia badań przed wystąpieniem oczekiwanego zdarzenia. Analiza przeżycia wyróżnia jednak wiele rodzajów cenzorowania, kategoryzując je według różnych kryteriów (Elektroniczny Podręcznik Statystyki, 2006):

- jedno- i wielokrotne – w zależności od tego, czy obserwacja wszystkich obiektów kończy się w tym samym momencie, czy też ucięcia są od siebie niezależne,

- typu I i II – obserwację kończy się w ustalonym momencie, lub po wystąpieniu badanego zdarzenia u założonej frakcji obiektów,

- prawo- i lewostronne – ucinanie może dotyczyć początku lub końca przedziału czasowego.

Formalnie, jeśli przez T oznaczymy czas do wystąpienia zdarzenia, a przez Z czas cenzorowania, to obserwacja[2] będzie uporządkowaną parą (X, δ), gdzie $X = \min(T, Z)$ a δ jest zmienną *indykatorową* $\delta = \lfloor T \leq Z \rceil$.[3]

Obserwacje ucięte – choć w pewnym sensie „niepełnowartościowe" – wnoszą jednak dodatkową wiedzę o badanym zjawisku i narzędzia stosowane w analizie przeżycia muszą ten fakt uwzględniać.

3.4 Nieparametryczna estymacja funkcji przeżycia

3.4.1 Estymator Kaplana-Meiera

Jednowymiarowa funkcja przeżycia może być estymowana za pomocą estymatora Kaplana-Meiera (1958), będącego estymatorem największej wiarogodności. Jego istotną cechą jest branie pod uwagę obserwacji cenzorowanych.

Niech $t_1 \leq t_2 \leq \ldots \leq t_n$ będzie n-elementową próbą uporządkowaną, n_i liczbą obiektów, u których zdarzenie nie wystąpiło przed czasem t_i, a d_i liczbą zdarzeń w momencie t_i. Estymator Kaplana-Meiera określony jest wzorem:

$$\bar{F}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}. \tag{22}$$

Odpowiednikiem estymatora Kaplana-Meiera dla skumulowanej funkcji hazardu jest estymator Nelsona-Aalena:

$$H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}. \tag{23}$$

[2] Dotyczy to przypadku cenzorowania prawostronnego.
[3] Zapis $\lfloor p \rceil$ oznacza funkcję przyjmującą wartość 1, kiedy zdanie p jest prawdziwe i wartość 0 w przeciwnym przypadku.

3.4.2 Estymacja dwuwymiarowej funkcji przeżycia

Sytuacja znacznie komplikuje się w przypadku dwuwymiarowym, kiedy każda ze zmiennych może być cenzorowana niezależnie – nie istnieje prosty odpowiednik estymatora Kaplana-Meiera.

Literatura przedmiotu zawiera szereg propozycji nieparametrycznych estymatorów dwuwymiarowych funkcji przeżycia. Dobrymi własnościami wyróżnia się estymator Dąbrowskiej (1988, 1989). Jest on jednak (podobnie jak inne propozycje) skomplikowany, przez co trudny w implementacji i wymagający sporych nakładów obliczeniowych.

Niech:

T_1, T_2 – dwuwymiarowa zmienna losowa,

Z_1, Z_2 – niezależne czasy cenzorowania,

X_1, X_2 – dwuwymiarowa zmienna losowa $X_i = \min(T_i, Z_i)$,

$\{(x_{1i}, x_{2i}, \delta_{1i}, \delta_{2i}),\ i = 1 \dots n\}$ - n-elementowa próba losowa,
gdzie $\delta_{ji} = \lfloor x_{ji} \leq z_{ji} \rceil$.

Dla uproszczenie i skrócenia zapisu przyjęto następującą konwencję:

$$f(\Delta s, t) = f(s, t) - f(s-, t), \tag{24}$$

$$f(s, \Delta t) = f(s, t) - f(s, t-), \tag{25}$$

$$f(\Delta s, \Delta t) = f(s, t) - f(s, t-) - f(s-, t) + f(s-, t-). \tag{26}$$

Niech H_{ij} będą pomocniczymi funkcjami przeżycia, odpowiednio:

$$H_{00}(s, t) = P(X_1 > s, X_2 > t), \tag{27}$$

$$H_{10}(s, t) = P(X_1 > s, X_2 > t, \delta_1 = 1), \tag{28}$$

$$H_{01}(s, t) = P(X_1 > s, X_2 > t, \delta_2 = 1), \tag{29}$$

$$H_{11}(s, t) = P(X_1 > s, X_2 > t, \delta_1 = 1, \delta_2 = 1). \tag{30}$$

Funkcje H_{ij} pozwalają na wyznaczenie skumulowanych funkcji hazardu:

$$\Lambda_{11}(s, t) = \int_0^s \int_0^t \frac{H_{11}(du, dv)}{H_{00}(u-, v-)}, \tag{31}$$

$$\Lambda_{10}(s, t) = -\int_0^s \frac{H_{10}(du, t)}{H_{00}(u-, t)}, \tag{32}$$

$$\Lambda_{01}(s, t) = -\int_0^t \frac{H_{01}(s, dv)}{H_{00}(s, v-)}. \tag{33}$$

Estymacja funkcji H_{ij} jest stosunkowo prosta:

$$\hat{H}_{00}(s,t) = \frac{1}{n}\sum_i \lfloor x_{1i} > s \ \wedge \ x_{2i} > t \rceil, \tag{34}$$

$$\hat{H}_{10}(s,t) = \frac{1}{n}\sum_i \lfloor x_{1i} > s \ \wedge \ x_{2i} > t \ \wedge \ \delta_1 = 1 \rceil, \tag{35}$$

$$\hat{H}_{01}(s,t) = \frac{1}{n}\sum_i \lfloor x_{1i} > s \ \wedge \ x_{2i} > t \ \wedge \ \delta_2 = 1 \rceil, \tag{36}$$

$$\hat{H}_{11}(s,t) = \frac{1}{n}\sum_i \lfloor x_{1i} > s \ \wedge \ x_{2i} > t \ \wedge \ \delta_1 = 1 \ \wedge \ \delta_2 = 1 \rceil, \tag{37}$$

co umożliwia z kolei estymację skumulowanych funkcji hazardu:

$$\hat{\Lambda}_{11}(s,t) = \int_0^s \int_0^t \frac{\hat{H}_{11}(du,dv)}{\hat{H}_{00}(u-,v-)}, \tag{38}$$

$$\hat{\Lambda}_{10}(s,t) = -\int_0^s \frac{\hat{H}_{10}(du,t)}{\hat{H}_{00}(u-,t)}, \tag{39}$$

$$\hat{\Lambda}_{01}(s,t) = -\int_0^t \frac{\hat{H}_{01}(s,dv)}{\hat{H}_{00}(s,v-)}. \tag{40}$$

Ostatecznie estymator dwuwymiarowej funkcji przeżycia wyznaczany jest z zależności:

$$\hat{F}(s,t) = \hat{F}(s,0)\cdot\hat{F}(0,t)\cdot\prod_{\substack{0<u\leq s\\ 0<v\leq t}}\left[1-\hat{L}(\Delta u,\Delta v)\right], \tag{41}$$

gdzie:

$$\hat{L}(\Delta u,\Delta v) = \frac{\hat{\Lambda}_{10}(\Delta u,v-)\hat{\Lambda}_{01}(u-,\Delta v)-\hat{\Lambda}_{11}(\Delta u,\Delta v)}{\left[1-\hat{\Lambda}_{10}(\Delta u,v-)\right]\left[1-\hat{\Lambda}_{01}(u-,\Delta v)\right]}. \tag{42}$$

$\hat{F}(s,0)$ i $\hat{F}(0,t)$ są jednowymiarowymi estymatorami Kaplana-Meiera:

$$\hat{F}(s,0) = \prod_{u\leq s}\left[1-\hat{\Lambda}_{10}(\Delta u,0)\right], \tag{43}$$

$$\hat{F}(0,t) = \prod_{v\leq t}\left[1-\hat{\Lambda}_{01}(0,\Delta v)\right]. \tag{44}$$

Dla danych niecenzorowanych estymator ten sprowadza się do zwykłej empirycznej funkcji przeżycia.

3.5 Kopuły w analizie przeżycia i niezawodności

Między dwuwymiarową funkcją przeżycia a jej jednowymiarowymi granicami zachodzi zależność analogiczna do tej, którą w terminach dystrybuant opisuje twierdzenie Sklara. Niech C będzie kopułą zmiennej losowej (X, Y). Łatwo wykazać, że:

$$\bar{F}(x, y) = \hat{C}(\bar{F}_x(x), \bar{F}_y(y)), \tag{45}$$

gdzie:

$$\hat{C}(u, v) = u + v - 1 + C(1 - u, 1 - v). \tag{46}$$

Warto w tym miejscu zwrócić szczególną uwagę na fakt – kluczowy dla problemów poruszanych w niniejszej pracy – że funkcja $\hat{C}$ jest *kopułą*, zatem dystrybuantą – nie funkcją przeżycia – zmiennej losowej o jednostajnych rozkładach brzegowych.

4 Modelowanie danych wielowymiarowych

Modelowanie wielowymiarowych danych przy użyciu kopuł jest zagadnieniem złożo-nym, nawet w przypadku ograniczenia obszaru zainteresowania do grupy kopuł archi-medesowskich z jednym parametrem.

Pierwszy z problemów to konieczność „jednoczesnego" wyboru rodziny kopuł i oszacowania wartości jej parametru. Oszacowanie wartości parametru kopuły nie jest możliwe bez znajomości jej postaci. Z drugiej strony, wybór dopasowanego do danych empirycznych modelu kopuły nie jest możliwy w oderwaniu od wartości jej parametru. Zagadnieniu temu poświęcony jest – w przeważającej części – ten rozdział.

Drugim problemem jest weryfikacja wyników modelowania. Może mieć ona dwojaką formę:

- potwierdzenie, że wybrany model lepiej od innych opisuje dane empiryczne,
- potwierdzenie, że wybrany model odpowiada prawdziwej strukturze zależności w badanej próbie losowej.

Propozycją odpowiedzi na pierwsze z tych pytań jest opisany w dalszej części rozdziału test istotności. Druga z kwestii wykracza poza tematykę niniejszej pracy.

Dodatkowe komplikacje pojawiają się w przypadku modelowania danych cenzorowanych. Najważniejszą z nich, poruszaną już w rozdziale 3.4, jest estymacja wielowymiarowej funkcji przeżycia.

4.1 Wybrane aspekty zastosowania kopuł w modelowaniu danych wielowymiarowych

Algorytm będący przedmiotem niniejszej pracy został zaprezentowany w artykule Chen, Fan (2007) i jest rozwinięciem koncepcji pojawiających się wcześniej w pracach Claytona (1978), Genesta i Rivesta (1993) oraz Wanga i Wellsa (2000).

Pierwsza z tych koncepcji, zaproponowana w pracy Genest, Rivest (1993), to wykorzystanie nieparametrycznego estymatora dystrybuanty kopuły $K(v)$ do wyznaczenia parametrów kopuły. Jest to możliwe dzięki udowodnionej przez autorów wzajemnie jednoznacznej relacji między kopułą archimedesowską a jej dystrybuantą:

$$K(v) = v - \frac{\phi(\theta)}{\phi'(\theta)}. \tag{47}$$

Ponadto autorzy wskazują na możliwość pominięcia etapu pośredniego jakim jest budowa dwuwymiarowej dystrybuanty empirycznej (formalnie niezbędnej do estymacji $K(v)$) oraz proponują oparcie estymatora parametru θ kopuły o dekompozycję współczynnika τ Kendalla.

Koncepcja powyższa znalazła rozwinięcie w pracy Wanga i Wellsa (2000). Jednak – inaczej niż poprzednicy – autorzy skoncentrowali się nad modelowaniem struktury zależności dla danych cenzorowanych, występujących w zagadnieniach analizy przeżycia i niezawodności.

Ważną konsekwencją takiego podejścia była konieczność wykorzystania nieparametrycznego estymatora dwuwymiarowej funkcji przeżycia do estymacji dystrybuanty kopuły $K(v)$.

Niech $x_1, \ldots, x_n$ oraz $y_1, \ldots, y_n$ będą uporządkowanymi obserwacjami z dwuwymiarowej próby losowej (x_i, y_i), $i = 1, \ldots, n$. Proponowany przez Wanga i Wellsa (2000) estymator $K(v)$ ma następującą postać:

$$\hat{K}(v) = 1 - \sum_{i=1}^{n}\sum_{y=1}^{n} \lfloor \hat{F}(x_i, y_i) > v \rceil \hat{F}(\Delta x_i, \Delta y_i), \tag{48}$$

gdzie $\widehat{F}$ jest nieparametrycznym estymatorem dwuwymiarowej funkcji przeżycia, natomiast:

$$\hat{F}(\Delta x_i, \Delta y_j) = \hat{F}(x_i, y_j) - \hat{F}(x_{i-1}, y_j) - \hat{F}(x_i, y_{j-1}) + \hat{F}(x_{i-1}, y_{j-1}). \tag{49}$$

Prezentowane podejście do estymacji $K(v)$ czyni tę procedurę bardziej elastyczną, pozwalając na dobór strategii estymacji dwuwymiarowej funkcji przeżycia odpowiednio do charakteru analizowanych danych losowych, w szczególności do rodzaju cenzorowania.

Następnie Wang i Wells (2000) proponują użycie $\hat{K}(v)$ do wyboru modelu najlepiej pasującego do analizowanych danych eksperymentalnych spośród arbitralnie wybranego zbioru rodzin kopuł archimedesowskich.

Podejście to wymaga wskazania jakiejś miary dopasowania poszczególnych modeli do danych losowych oraz kryterium wyboru modelu dopasowanego najlepiej. Miarą zaproponowaną w omawianej pracy jest scałkowany kwadrat różnicy między teoretyczną wartością dystrybuanty kopuły a jej wyestymowanym odpowiednikiem:

$$S(\alpha) = \int_{\xi}^{1} \left[\hat{K}(v) - K_{\alpha}(v)\right]^2 dv, \tag{50}$$

zaś metoda wyboru najlepiej dopasowanego modelu sprowadza się do wyboru kopuły, dla której $S(\alpha)$ ma wartość najmniejszą.

Wyznaczenie $K_{\alpha}(v)$ (a co za tym idzie $S(\alpha)$) dla konkretnego modelu wymaga estymacji parametru α, która może bazować na współczynniku τ Kendalla, dzięki przedstawionej w rozdziale 2.5 zależności:

$$\tau = 4 \int_{0}^{1} \frac{\phi_{\alpha}(v)}{\phi'_{\alpha}(v)} dv + 1. \tag{51}$$

Zależność ta pozwala wprawdzie na stosunkowo proste wyznaczenie parametru α na postawie wartości τ, zasadniczym problemem jest jednak nieparametryczna estymacja współczynnika τ dla danych cenzorowanych.

4.2 Algorytm wyboru modelu i estymacji parametrów kopuły

Przestawione w rozdziale 4.1 koncepcje zostały usystematyzowane i rozszerzone w pracy Chen, Fan (2007). Położono w niej nacisk na praktyczne aspekty zastosowania przedstawionej w niej procedury. Przejawia się to – między innymi – konsekwentnie stosowanym i podkreślanym w pracy założeniem, że żadna z rodzin kopuł – kandydatów może nie opisywać prawdziwej zależności w rozkładzie, z którego pochodzi próba losowa.

Jest to podejście bliskie zagadnieniom praktycznym, w których konieczne jest wybranie, spośród ograniczonego zbioru dostępnych modeli, jednego – najlepiej odpowiadającego danym eksperymentalnym.

Niech:

T_x, T_y – dwuwymiarowa zmienna losowa,

Z_x, Z_y - niezależne czasy cenzorowania,

X, Y – dwuwymiarowa zmienna losowa: $X = \min(T_x, Z_x), Y = \min(T_y, Z_y)$,

$\{(x_i, y_i, \delta_{xi}, \delta_{yi}),\ i = 1 \ldots n\}$ - n-elementowa próba losowa, gdzie $\delta_{xi} = \lfloor x_i \leq z_{xi} \rceil$, $\delta_{yi} = \lfloor y_i \leq z_{yi} \rceil$,

$\{C_j(u, v, \alpha_j),\ j = 1 \ldots m\}$ – m-elementowy zbiór kopuł archimedesowskich – potencjalnych modeli,

$\{K_j(v, \alpha_j),\ j = 1 \ldots m\}$ – dystrybuanty kopuł $\{C_j\}$.

Procedura wyboru modelu i estymacji jego parametrów ma następujący przebieg:

1. nieparametryczna estymacja dwuwymiarowej funkcji przeżycia $\bar{F}(x, y)$,
2. nieparametryczna estymacja dystrybuanty kopuły $K(v)$,
3. estymacja parametru α dla każdej kopuły ze zbioru modeli kandydujących,
4. wybór modelu najlepiej dopasowanego do danych losowych,
5. test istotności dla kryterium wyboru modelu.

Punkty 1 i 2 nie różnią się od omawianych już propozycji zawartych w pracy Wanga i Wellsa (2000). Estymator dwuwymiarowej funkcji przeżycia $\bar{F}(x, y)$ jest jednym z parametrów procedury i jego konkretna postać, zależna jest jedynie od własności próby losowej.

Estymacja dystrybuanty kopuły $K(v)$ odbywa się na podstawie znanej już zależności:

$$\hat{K}(v) = 1 - \sum_{i=1}^{n} \sum_{j=1}^{n} \lfloor \hat{F}(x_i, y_j) > v \rceil \cdot \hat{F}(\Delta x_i, \Delta y_j), \tag{52}$$

gdzie: $x_1 \leq x_2 \leq \ldots x_n$ oraz $y_1 \leq y_2 \leq \ldots y_n$ są uporządkowanymi obserwacjami z próby losowej $\{(x_i, y_i),\ i = 1 \ldots n\}$, oraz:

$$\hat{F}(\Delta x_i, \Delta y_j) = \hat{F}(x_i, y_j) - \hat{F}(x_{i-1}, y_j) - \hat{F}(x_i, y_{j-1}) + \hat{F}(x_{i-1}, y_{j-1}). \tag{53}$$

Innowacją w stosunku do omawianych wcześniej metod jest estymacja parametrów poszczególnych kopuł oparta o minimalizację scałkowanego kwadratu różnicy:

$$\hat{S}_j(\alpha_j) = \int_{\xi}^{1} \left[\hat{K}(v) - K_j(v, \alpha_j) \right]^2 dv, \tag{54}$$

która w zastosowaniach praktycznych może być zastąpiona sumą Riemanna (Wang, Wells, 2000):

$$\hat{S}_j(\alpha_j) = \sum_{i=1}^{n} \left[\hat{K}(v_i) - K_j(v_i, \alpha_j) \right]^2 \cdot (v_i - v_{i-1}), \tag{55}$$

gdzie: $v_0 = 0$, natomiast $v_1 \leq v_2 \leq \ldots \leq v_n$ to uporządkowany zbiór wartości

$$\{v_i = \hat{F}(x_i, y_i),\ i = 1 \ldots n\}.$$

Dla każdej z kopuł estymatorem parametru α jest wartość, dla której scałkowany kwadrat różnicy osiąga minimum:

$$\hat{\alpha}_j = \arg\min_{\alpha_j} \int_{\xi}^{1} \left[\hat{K}(v) - K_j(v, \alpha_j)\right]^2 dv = \arg\min_{\alpha_j} \hat{S}_j(\alpha_j). \tag{56}$$

Należy zwrócić w tym miejscu uwagę na fakt, że nie wszystkie modele $\{C_j\}$ opisują prawdziwą formę zależności w rozkładzie, z którego pochodzi próba losowa (w szczególnym, choć nie rzadkim w praktyce, przypadku – żaden z nich). Zatem formalnie $\hat{\alpha}_j$ nie jest estymatorem prawdziwej wartości α_j a jedynie takiej wartości parametru – α_j^*, która najlepiej dopasowuje dany model do analizowanych danych.

Kryterium wyboru najlepiej dopasowanego modelu jest wartość $\hat{S}_j(\hat{\alpha}_j)$. Ze zbioru kopuł $\{C_j\}$ wybierana jest ta, dla której $\hat{S}_j(\hat{\alpha}_j)$ osiąga wartość najmniejszą.

4.3 Test istotności dla kryterium wyboru modelu

Ostatnim elementem procedury proponowanej w pracy Chen, Fan (2007), nie mającym odpowiednika w innych z omawianych prac jest statystyczny test istotności dla kryterium wyboru modelu.

Wyjaśnienie jego idei ułatwia rozważenie przypadku, gdy zbiór dostępnych modeli – kopuł $\{C_j(u, v, \alpha)\}$ jest dwuelementowy, a najmniejszą wartość $\hat{S}_j(\hat{\alpha}_j)$ uzyskano dla modelu C_1. Hipoteza zerowa i hipoteza alternatywna byłyby w takim przypadku następujące:

$$H_0 : \; S_1(\alpha_1^*) \leq S_2(\alpha_2^*), \qquad H_1 : \; S_1(\alpha_1^*) > S_2(\alpha_2^*).$$

Odrzucenie hipotezy zerowej skutkowałoby wybraniem modelu C_2, w przypadku braku podstaw do jej odrzucenia – wybrany zostałby model C_1.

Hipoteza zerowa i alternatywna wymagają nieco innej konstrukcji w przypadku, kiedy zbiór rozpatrywanych modeli ma więcej elementów. Podobnie jak poprzednio założono, że modelem odniesienia jest kopuła C_1:

$$H_0 : \max_{j=2,\ldots,m} \left[S_1(\alpha_1^*) - S_j(\alpha_j^*)\right] \leq 0, \qquad H_1 : \max_{j=2,\ldots,m} \left[S_1(\alpha_1^*) - S_j(\alpha_j^*)\right] > 0.$$

Tym razem, w przypadku odrzucenia hipotezy zerowej wybrany zostałby model ze zbioru $\{C_j,\ j = 2, \ldots, m\}$, dla którego wartość $\hat{S}_j(\hat{\alpha}_j)$ byłaby najmniejsza. Jak podkreślają autorzy pracy, tak sformułowany test nie wymaga, aby którakolwiek z kopuł C_j odpowiadała prawdziwej kopule w badanym rozkładzie, również w przypadku przyjęcia hipotezy zerowej.

Statystyką testową w tym teście jest:

$$T_n = \max_{j=2,\ldots,m} \sqrt{n} T_{jn} = \max_{j=2,\ldots,m} \sqrt{n} \left[\hat{S}_1(\hat{\alpha}_1) - \hat{S}_j(\hat{\alpha}_j)\right], \tag{57}$$

a kryterium odrzucenia hipotezy H_0:

$$T_n > Z_\alpha, \tag{58}$$

gdzie Z_α jest górnym α-percentylem rozkładu statystyki testowej. Rozkład ten – w ogólnym przypadku – jest nieznany.

Autorzy proponują zastosowanie procedury „naiwnego boostrapu" do aproksymacji nieznanego rozkładu statystyki testowej:

1. niech $\{(x_i^*, y_i^*, \delta_{xi}^*, \delta_{yi}^*),\ i = 1, \ldots, n\}$ będzie próbą losowaną ze zwracaniem z oryginalnej próby losowej $\{(x_i, y_i, \delta_{xi}, \delta_{yi}),\ i = 1, \ldots, n\}$ a $\hat{F}^*$, $\hat{K}^*$ i $\hat{\alpha}_j^*$ będą wyznaczonymi na podstawie tej próby odpowiednikami $\hat{F}$, $\hat{K}$ oraz $\hat{\alpha}_j$,

2. niech $T_n^* = \max_{j=2,\ldots,m} \sqrt{n}\left(T_{jn}^* - T_{jn}\right)$, gdzie T_{jn}^* jest odpowiednikiem T_{jn} wyznaczonym na podstawie próby bootstrap,

3. kroki 1-2 należy powtórzyć wielokrotnie i użyć dystrybuanty empirycznej uzyskanych wartości T_n^* do aproksymacji rozkładu statystyki testowej.

5 Implementacja algorytmu

Metoda modelowania danych losowych, omówiona w poprzednim rozdziale, została zaimplementowana w postaci aplikacji **biFail**, napisanej w języku Java. Z uwagi na dość skomplikowany algorytm modelowania, celem jaki postawiono było stworzenie poprawnie działającego prototypu, a nie aplikacji przeznaczonej dla użytkownika końcowego.

Jedyną biblioteką zewnętrzną wykorzystaną w aplikacji **biFail** jest biblioteka Apache Math Commons[4]. Aplikacja używa udostępnianej przez nią implementacji algorytmu Brenta, przeznaczonego do poszukiwania minimów funkcji rzeczywistych (Brent, 1973). Algorytm ten zastosowano do estymacji parametru α kopuły na podstawie scałkowanego kwadratu różnicy między teoretyczną dystrybuantą kopuły a jej empirycznym odpowiednikiem.

6 Eksperymenty numeryczne

6.1 Modelowanie danych

Eksperymenty numeryczne przeprowadzono na sztucznie wygenerowanych danych losowych pochodzących z rozkładów o ustalonych funkcjach przeżycia. Podzielono je na trzy etapy.

[4] http://commons.apache.org/math/

Najpierw przetestowano działanie algorytmu dla danych niecenzorowanych, następnie dla danych cenzorowanych o jednostajnych rozkładach brzegowych. Ostatni etap eksperymentów dotyczył danych cenzorowanych o rozkładach brzegowych normalnych.

Każdy z etapów składał się z trzech eksperymentów:

1. kopuła Claytona, $\alpha = 0.86$ ($\tau = 0.3$),

2. kopuła Claytona, $\alpha = 4.67$ ($\tau = 0.7$),

3. kopuła Gumbela, $\alpha = 2.33$ ($\tau = 0.7$).

W każdym z eksperymentów przeprowadzono analizę 100 losowych próbek, pochodzących z identycznych rozkładów, składających się z 200 obserwacji.

Zbiór kopuł archimedesowskich używanych w charakterze modeli składał się z kopuł Claytona, Franka i Gumbela. Na podstawie analizy wykresów rozrzutu obszar poszukiwań dla $\hat{\alpha}$ ograniczono, identycznie dla każdego z modeli, do przedziału $\langle 1, 50 \rangle$.

6.1.1 Dane niecenzorowane

Wyniki doświadczeń z danymi niecenzorowanymi przedstawiono w tabeli 1.

W przypadku kopuły Claytona o parametrze $\alpha = 0.86$ model został niepoprawnie rozpoznany w 20 przypadkach i w każdym z tych przypadków wybierana była kopuła Franka. Wyjaśnieniem tego zjawiska może być niewielki poziom zależności w analizowanych danych ($\tau = 0.3$). Wraz ze spadkiem wartości bezwzględnej tego współczynnika różnice pomiędzy poszczególnymi kopułami zaczynają się zacierać.

Średnie wartości $\hat{\alpha}$ i $\hat{S}(\hat{\alpha})$ wyznaczono na podstawie tylko tych przypadków, w których kopuła została zidentyfikowana prawidłowo. Tak postępowano również we wszystkich kolejnych doświadczeniach.

Wyniki testu istotności nie dały podstaw do odrzucenia hipotezy zerowej w żadnym z przypadków (również w tych, w których wybrany został błędny model). Minimalna wartość granicznego poziomu istotności p-value wyniosła 0.75 dla kopuły Claytona i 0.5 dla kopuły Franka.

Liczba nieprawidłowo rozpoznanych modeli zmalała do jednego w przypadku kopuły Claytona o parametrze $\alpha = 4.67$. W jednym przypadku najlepiej dopasowanym modelem okazała się ponownie kopuła Franka. Wzrosła też dokładność oszacowania parametru kopuły. Tak jak poprzednio w żadnej z analiz nie odrzucono hipotezy zerowej, a minimalna p-value wyniosła 0.87.

Ostatnią kopułą była kopuła Gumbela ($\alpha = 2.33$). W czterech przypadkach została rozpoznana jako kopuła Franka. Minimalna wartość p-value wyniosła 0.57.

6.1.2 Jednostajne rozkłady brzegowe

Drugi etap eksperymentów przeprowadzono na danych pochodzących z rozkładów identycznych jak w etapie pierwszym, ale każda z próbek była cenzorowana (niezależnie) rozkładem Weibulla ($k = 3$, $\lambda = 1.2$). Przy doborze parametrów rozkładu Weibulla kierowano się przede wszystkim założonym odsetkiem obserwacji cenzorowanych wynoszącym ok. 0.25.

W przypadku kopuły Claytona ($\alpha = 0.86$) i kopuły Gumbela liczba przypadków błędnej identyfikacji modelu wzrosła. Kopuła Claytona ($\alpha = 4.67$) została rozpoznana bezbłędnie.

Podobnie jak w przypadku danych niecenzorowanych, nie odnotowano odrzucenia hipotezy zerowej, a minimalne wartości granicznego poziomu istotności nie odbiegały od wyników uzyskanych w poprzednim etapie.

kopuła	Clayton		Gumbel
α	0.86	4.67	2.33
τ	0.3	0.7	0.7
błędne rozpoznanie	20	1	4
średnia $\hat{\alpha}$	1.022	4.689	2.311
średnia $\hat{S}(\hat{\alpha})$	$3.727 \cdot 10^{-4}$	$1.676 \cdot 10^{-4}$	$2.076 \cdot 10^{-4}$
odrzucenie H_0	0	0	0

Table 1: Wyniki eksperymentów dla danych niecenzorowanych

kopuła	Clayton		Gumbel
α	0.86	4.67	2.33
τ	0.3	0.7	0.7
częstość cenzorowania	0.253	0.243	0.25
błędne rozpoznanie	26	0	7
średnia $\hat{\alpha}$	1.06	4.487	2.305
średnia $\hat{S}(\hat{\alpha})$	$3.746 \cdot 10^{-4}$	$2.617 \cdot 10^{-4}$	$2.684 \cdot 10^{-4}$
odrzucenie H_0	0	0	0

Table 2: Wyniki eksperymentów dla danych cenzorowanych o jednostajnych rozkładach brzegowych

6.1.3 Normalne rozkłady brzegowe

Ostatnim etapem eksperymentów była analiza danych o rozkładach brzegowych normalnych $N(0.5, 0.25)$. Parametry rozkładów brzegowych zostały dobrane tak, aby utrzymać częstość cenzorowania na poziomie podobnym jak w etapie poprzednim (parametry kopuł i rozkładu Weibulla pozostały bez zmian).

Niejednostajne rozkłady brzegowe wpłynęły na prawie dwukrotny wzrost błędnych identyfikacji w przypadku kopuły Claytona o słabszej zależności i kopuły Gumbela (choć tutaj ich liczba nadal pozostawała niewielka). Kopuła Claytona o silniejszej zależności, podobnie jak w przypadku eksperymentów z danymi niecenzorowanymi, nie została rozpoznana tylko raz. W tym przypadku spadła dość wyraźnie dokładność oszacowania parametru kopuły.

Ponownie nie zaobserwowano wartości p-value dającej podstawy do odrzucenia hipotezy zerowej.

kopuła	Clayton		Gumbel
α	0.86	4.67	2.33
τ	0.3	0.7	0.7
częstość cenzorowania	0.215	0.22	0.215
błędne rozpoznanie	42	1	14
średnia $\hat{\alpha}$	1.018	4.099	2.204
średnia $\hat{S}(\hat{\alpha})$	$4.59 \cdot 10^{-4}$	$2.371 \cdot 10^{-4}$	$2.494 \cdot 10^{-4}$
odrzucenie H_0	0	0	0

Table 3: Wyniki eksperymentów dla danych cenzorowanych o rozkładach brzegowych normalnych

6.2 Wnioski

Przeprowadzone eksperymenty numeryczne potwierdziły poprawne działanie algorytmu modelowania. Pokazały też (możliwe do przewidzenia) zmniejszanie się jego efektywności dla danych o słabej zależności. Wyraźny wpływ na działanie algorytmu mają też rozkłady brzegowe prób, choć wpływ ten słabnie przy danych o silniejszych zależnościach.

Nie udało się zaobserwować przypadku odrzucenia hipotezy zerowej. Wynik ten nie jest jednak zaskakujący dla sztucznie generowanych zestawów danych.

Bibliografia

Armstrong M. (2003) Copula catalogue, part 1: Bivariate archimedean copulas, dostępne (14.09.2015) na stronie http://www.researchgate.net/publication/228919 559_Copula_ catalogue._Part_1_bivariate_Archimedean_copulas.

Brent R. P. (1973) *Algorithms for Minimization without Derivatives*, Prentice-Hall, Englewood Cliffs, N.J.

Chen X., Fan Y. (2007) A model selection test for bivariate failure-time data,*Econometric Theory*, 23, 414–439,.

Clayton D. G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, 65, 141–151.

Dąbrowska D. M. (1988) Kaplan-Meier estimate on the plane, *The Annals of Statistics*, 16: 1475–1489.

Dąbrowska D. M. (1989) Kaplan-Meier estimate on the plane: Weak convergence, LIL and the bootstrap, *Journal of Multivariate Analysis*, 29, 308–325.

Elektroniczny Podręcznik Statystyki PL (2006) StatSoft, Kraków, http://www.statsoft. pl/text book/stathome.html.

Frees E. W., Valdez E. A. (1998) Understanding relationships using copulas, *North American Actuarial Journal*, 2(1), 1–25.

Genest Ch., MacKay J. (1986) The joy of copulas: Bivariate distributions with uniform marginals, *The American Statistician*, 40, 280–283.

Genest Ch., Rivest L. P. (1993) Statistical inference procedures for bivariate archimedean copulas, *Journal of the American Statistical Association*, 88, 1034–1043.

Kaplan E. L., Meier P. (1958) Nonparametric estimation from in complete observations, *Journal of the American Statistical Association*, 53, 457–481.

Kendall M. G. (1948) *Rank correlation methods*, Griffin, London.

Sklar A. (1959) *Fonctions de répartition á n dimensions et leurs marges*, Inst. Statist. Univ. Paris, Paris.

Wang W., Wells M. T. (2000) Model selection and semiparametric inference for bivariate failure-time data, *Journal of the American Statistical Association*, 95, 62–76.

COMPUTER MODELLING OF MULTIVARIATE DATA USING COPULAS

Abstract: The subject of the paper is constituted by the implementation of the method of modelling of the censored multidimensional data with the use of the copula, proposed in the paper by Wang and Wells (2000).

The paper provides the basic information on the copulas, the issues related to survival and reliability questions, as well as consideration of selected aspects of modelling of the multidimensional data. Then, details are presented of the implementation of a method for modelling of censored data, along with the results from the numerical experiments, carried out with the artificially generated random data.

Keywords: copulas, multidimensional data, censored data, modelling, reliability