Drzewa, błądzenia i transformata Fouriera

Jacek Jendrej

14 lipca 2012

Streszczenie

Tematem eseju jest zastosowanie transformaty Fouriera w filogenetyce. Omawiamy proste uogólnienie klasycznych reguł rachunku funkcji charakterystycznych w teorii prawdopodobieństwa, które może być zastosowane do badania grupowych modeli ewolucji.

1 Wprowadzenie

Zadanie filogenetyki można opisać następująco. Danych jest zbiór $n-1$ gatunków $L' = \{L_1, \ldots, L_{n-1}\}$. Zakładamy, że dla każdego z tych gatunków rozkodowano fragment DNA długości N, przy czym ustalono, że ten fragment pochodzi od tego samego odcinka DNA ich wspólnego przodka R (tzn. wykonano *multiple sequence alignment*). W ten sposób otrzymujemy N ciągów długości $n-1$

$$g^l = (g^l_1, \ldots, g^l_{n-1}), \qquad l = 1, \ldots, N,$$

gdzie $g^l_i \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ jest jednym z czterech nukleotydów. Na podstawie g^l należy określić drzewo ewolucji gatunków $L_1, \ldots, L_{n-1}$ od przodka R.

W tym celu zakłada się, że mutacje DNA są procesem losowym, który przebiega niezależnie na poszczególnych pozycjach łańcucha DNA. Dla przykładu dwuparametrowy model Kimury [2] $K(\alpha, \beta)$ przyjmuje, że pojedynczy „krok ewolucyjny" polega na wykonaniu tranzycji z prawdopodobieństwem α, natomiast transwersji z prawdopodobieństwem β (tranzycja to zamiana $\mathrm{A} \leftrightarrow \mathrm{G}$ lub $\mathrm{C} \leftrightarrow \mathrm{T}$; transwersje to pozostałe podstawienia, które są znacznie mniej prawdopodobne).

Ciąg $(g^1, \ldots, g^N)$ stanowi próbę losową, która wyznacza rozkład empiryczny na zbiorze $\{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}^{L'}$. Przyjmuje się, że wiarygodnym drzewem ewolucji jest takie drzewo, dla którego rozkład teoretyczny otrzymany z modelu jest w jakimś sensie zbliżony do rozkładu empirycznego. W niniejszym eseju zajmujemy się problemem znalezienia rozkładu teoretycznego na liściach.

Przedstawiony materiał oparty jest na referacie wygłoszonym przez M. Zdanowicza na seminarium „Algebraiczne modele drzew filogenetycznych" prowadzonym przez prof. J. Wiśniewskiego. Głównym źródłem jest [7]. Dowody przytaczanych faktów z teorii prawdopodobieństwa można znaleźć w [6]. Omówienie analizy Fouriera na skończonych grupach abelowych zawiera np. [3, Wykład 104]. Wprowadzenia do filogenetyki z punktu widzenia biologa dostarcza [5].

2 Zmienne losowe o wartościach w skończonej grupie abelowej

Niech G będzie ustaloną skończoną grupą abelową. Przez $\widehat{G}$ oznaczamy grupę charakterów grupy G. Jest to grupa abelowa z działaniem grupowym $(\chi_1 + \chi_2)(g) = \chi_1(g)\chi_2(g)$ (po prawej mamy *mnożenie* w ciele $\mathbb{C}$).

Jeśli $\chi_1, \ldots, \chi_k \in \widehat{G}$, to przez $(\chi_1, \ldots, \chi_k)$ rozumiemy charakter χ grupy G^k zadany przez $\chi(g_1, \ldots, g_k) = \prod_{i=1}^{k} \chi_i(g_i)$. Łatwo widać, że wszystkie charaktery grupy G^k są tej postaci, tzn. grupą charakterów grupy G^k jest $\widehat{G}^k$.

Zmienna losowa o wartościach w G to funkcja mierzalna $X : \Omega \to G$, gdzie $(\Omega, \mathscr{F}, \mathbb{P})$ jest przestrzenią probabilistyczną. Zbiór zmiennych losowych o wartościach w G jest $\mathbb{Z}$-modułem z dodawaniem i mnożeniem przez skalary zdefiniowanym po wartościach. Mówimy, że zmienne losowe $X_1, X_2, \ldots, X_k$ o wartościach w G są *niezależne*, jeśli dla dowolnych $g_1, \ldots, g_k \in G$ zdarzenia $\{X_1 = g_1\}, \ldots, \{X_k = g_k\}$ są niezależne. Zauważmy, że wartość oczekiwana zmiennej losowej o wartościach w G nie jest zdefiniowana. Natomiast dla ustalonego charakteru $\chi \in \widehat{G}$ funkcja $\chi \circ X$ jest zmienną losową o wartościach w $\mathbb{C}$ i ma dobrze zdefiniowaną wartość oczekiwaną $\mathbb{E}(\chi \circ X)$.

Ciąg zmiennych losowych $X_1, \ldots, X_k$ o wartościach w G definiuje zmienną losową $X = (X_1, \ldots, X_k)^T$ o wartościach w G^k (przyjmujemy konwencję, że wektory losowe są wektorami kolumnowymi, natomiast charaktery na grupie G^k zapisujemy jako wektory wierszowe).

Definicja 2.1. Niech X będzie zmienną losową o wartościach w G. *Funkcją charakterystyczną* zmiennej X nazywamy funkcję $\varphi_X : \widehat{G} \to \mathbb{C}$ zadaną wzorem

$$\varphi_X(\chi) = \mathbb{E}(\chi \circ X).$$

Poniższe twierdzenie orzeka, że funkcja charakterystyczna jednoznacznie określa rozkład zmiennej losowej.

Twierdzenie 2.2. *Niech X będzie zmienną losową o wartościach w G i niech $g \in G$. Wówczas*

$$\mathbb{P}(X = g) = \frac{1}{|G|} \sum_{\chi \in \widehat{G}} \overline{\chi(g)} \varphi_X(\chi). \tag{1}$$

Dowód. Jest to wzór na odwrotną transformatę Fouriera. □

Twierdzenie 2.3. *Niech $X_1, \ldots, X_k$ będą niezależnymi zmiennymi losowymi o wartościach w G, niech $X = (X_1, \ldots, X_k)^T$ i $\chi = (\chi_1, \ldots, \chi_k)$. Wówczas*

$$\varphi_X(\chi) = \prod_{i=1}^{k} \varphi_{X_i}(\chi_i).$$

Odpowiednik Twierdzenia 2.3 w przypadku $G = \mathbb{R}$ jest podstawowym faktem w teorii prawdopodobieństwa. Dowód przenosi się na przypadek skończonej grupy abelowej bez istotnych modyfikacji.

Zauważmy, że jeśli $g = (g_1, \ldots, g_p)^T \in G^p$ i $A = (a_{ji})$ jest macierzą rozmiaru $q \times p$ o współczynnikach całkowitych, to wzór $h = Ag$ definiuje element $h \in G^q$. Ponadto dla $\psi = (\psi_1, \ldots, \psi_q) \in \widehat{G}^q$ wzór $\chi = \psi A$ określa charakter $\chi \in \widehat{G}^p$. Zachodzi przy tym prawo łączności

$$\psi \circ Ag = \prod_{i=1}^{p} \prod_{j=1}^{q} \psi_j(g_i)^{a_{ji}} = \psi A \circ g.$$

Wniosek 2.4. *Niech $X_1, \ldots, X_p$ będą niezależnymi zmiennymi losowymi o wartościach w G i niech $A \in \mathscr{M}_{q \times p}(\mathbb{Z})$. Wówczas funkcja charakterystyczna zmiennej losowej $Y = A(X_1, \ldots, X_p)^T$ dana jest wzorem*

$$\varphi_Y(\psi) = \prod_{i=1}^{p} \varphi_{X_i}\left((\psi A)_i\right).$$

Dowód. Z definicji mamy $\varphi_Y(\psi) = \mathbb{E}(\psi \circ Y) = \mathbb{E}(\psi \circ AX) = \mathbb{E}(\psi A \circ X) = \varphi_X(\psi A)$. Powołanie się na Twierdzenie 2.3 kończy dowód. □

Wzór ten możemy traktować jako uogólnienie Twierdzenia 2.3, które otrzymujemy dla $A = \mathrm{Id}$. Odnotujmy też, że dla $A = (1, \ldots, 1)$ dostajemy klasyczny wzór na funkcję charakterystyczną sumy niezależnych zmiennych losowych

$$\varphi_{X_1+\ldots+X_k} = \prod_{i=1}^{k} \varphi_{X_i}. \tag{2}$$

3 Błądzenia i ich transformaty

Definicja 3.1. Niech $X_1, \ldots, X_p$ będzie ciągiem niezależnych zmiennych losowych o wartościach w G. Niech $S_k = \sum_{i=1}^{k} X_i$. Wówczas ciąg zmiennych losowych $(S_0, S_1, \ldots, S_p)$ nazywamy (skończonym) *błądzeniem na grupie G*.

Uwaga 3.2. Wzór (2) pozwala liczyć funkcje charakterystyczne błądzeń.

Błądzenia modelują procesy „liniowe" składające się z pewnej liczby kroków, z których każdy polega na dodaniu pewnej losowej wartości do naszego „stanu konta". Takim procesem jest na przykład seria rzutów monetą, jeśli zliczamy ile razy wypadł orzeł ($G = \mathbb{Z}$). Ostatnia cyfra w zapisie dziesiętnym liczby orłów jest przykładem błądzenia na skończonej grupie $G = \mathbb{Z}_{10}$.

Okazuje się, że niektóre modele ewolucji dopuszczają podobną interpretację. Rozważmy na przykład model Kimury opisany we Wprowadzeniu. Potraktujmy $G = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ jako grupę $\mathbb{Z}_2 \times \mathbb{Z}_2$ z jednością A (oczywiście równie dobrze mógłby to być dowolny inny nukleotyd). Wówczas łatwo się przekonać, że dodanie G odpowiada tranzycji, zaś dodanie C lub T daje transwersję. W ten sposób pojedynczy krok ewolucji według dwuparametrowego modelu Kimury $K(\alpha, \beta)$ jest równoważny dodaniu zmiennej losowej X o wartościach w G i rozkładzie $(\mathbb{P}(X = \mathrm{A}), \mathbb{P}(X = \mathrm{G}), \mathbb{P}(X = \mathrm{C}), \mathbb{P}(X = \mathrm{T})) = (1 - \alpha - 2\beta, \alpha, \beta, \beta)$.

Aby opisać możliwość rozgałęziania się procesu ewolucji, uogólnimy pojęcie błądzenia, definiując błądzenia na drzewach. Symbol T oznaczać będzie drzewo, V zbiór wierzchołków, L zbiór liści, E zbiór krawędzi, a $R \in L$ korzeń. Dla wygody oznaczamy $L' = L \setminus \{R\}$. Dla $v \in V$ przez π_v oznaczamy zbiór krawędzi na ścieżce od R do v. Dla $e \in E$ definiujemy $L_e = \{l \in L : e \in \pi_l\}$.

Definicja 3.3. Niech dane będzie drzewo T. Niech dla każdej krawędzi $e \in E$ dana będzie zmienna losowa X_e, przy czym wszystkie te zmienne są niezależne. Dla $v \in V$ przyjmijmy $S_v = \sum_{e \in \pi_v} X_e$. Wówczas ciąg zmiennych losowych $(S_v)_{v \in V}$ nazywamy *T-błądzeniem na grupie G*.

Niech drzewo T ma n liści, tzn. $\left|L'\right| = n - 1$. Odpowiednikiem zmiennej S_p z błądzenia klasycznego jest w przypadku T-błądzenia zmienna losowa o wartościach w G^{n-1} zdefiniowana jako $S = (S_l)_{l \in L'}^{T}$. Oznaczmy $X = (X_e)_{e \in E}^{T}$ i określmy macierz $A = (a_{l,e}) \in \mathcal{M}_{|L'| \times |E|}(\mathbb{Z})$ za pomocą reguły

$$a_{l,e} = \begin{cases} 1 & \text{jeśli } e \in \pi_l, \\ 0 & \text{w przeciwnym wypadku.} \end{cases}$$

Wówczas $S = AX$, więc Wniosek 2.4 pozwala policzyć funkcję charakterystyczną zmiennej losowej S, jeśli znane są funkcje charakterystyczne zmiennych losowych X_e. Uzyskany wzór możemy zapisać w postaci

$$\varphi_S(\psi) = \varphi_X\left((\chi_e)_{e \in E}\right), \tag{3}$$

gdzie $\psi = (\psi_l)_{l \in L'}$, $\chi_e = \sum_{l \in L_e} \psi_l$.

4 Końcowe uwagi

Wzór (3) w prosty sposób wiąże przyjęty model ewolucji z rozkładem teoretycznym na liściach. Funkcje charakterystyczne zmiennych losowych X_e są bezpośrednio dane przez model, natomiast reguła obliczania charakterów χ_e przez topologię drzewa.

Można pójść dalej. Udowodniono pewne tożsamości kombinatoryczne, które pozwalają przekształcić prawą stronę równości (3). Uzyskuje się w ten sposób wzory, które umożliwiają wyznaczenie rozkładu zmiennych X_e na podstawie rozkładu zmiennej S (zob. [7, Rozdział 3]). Dla $G = \mathbb{Z}_2$ rezultat ten uzyskał już Hendy [1]. Twierdzenie Hendy'ego było z powodzeniem wykorzystywane w rzeczywistych obliczeniach (por. [4, Rozdział 8.6]).

Literatura

[1] M. D. Hendy. The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology*, (38):310–321, 1989.

[2] M. Kimura. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences, USA*, (78):454–458, 1981.

[3] T. W. Körner. *Fourier Analysis*. Cambridge University Press, 1988.

[4] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.

[5] S. C. Stearns. *Phylogeny and Systematics*. Yale University: Open Yale Courses. `http://oyc.yale.edu/ecology-and-evolutionary-biology/eeb-122/lecture-15`.

[6] K. R. Stromberg. *Probability for Analysts*. Chapman & Hall, 1994.

[7] L. A Székely, M. Steel, and P. L. Erdős. Fourier calculus on evolutionary trees. *Advances in Appl. Math.*, (14):200–216, 1993.