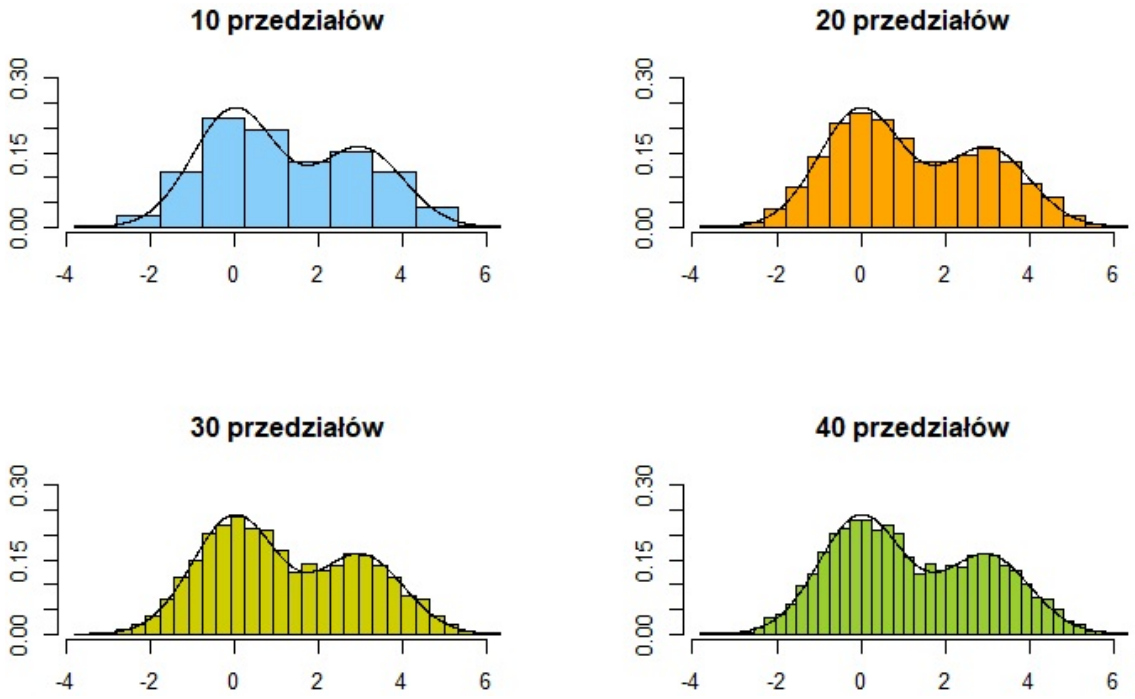
Estymacja jądrowa gęstości

Szymon Meyer

28.04.2021

1 Wprowadzenie

Rozważmy na początku jednowymiarową zmienną losową o rozkładzie opisanym poprzez nieznaną gęstość f. Niech x_1, x_2, ..., x_n będzie próbą z tego rozkładu. Często używanym estymatorem f jest histogram.



Rysunek 1: Przykładowe histogramy z różną liczbą klas. Ciągłym czarnym wykresem pokazano prawdziwą gęstość rozkładu.

Zauważmy, że dla każdej realizacji próby, histogram jest gęstością pewnego rozkładu prawdopodobieństwa. Jak widzimy na Rysunku 1, gdy rośnie liczba klas histogramu, to zwiększa się także dokładność przybliżenia gęstości rozkładu. Niestety, w żadnym przypadku histogram nie jest funkcją ciągłą, co można uznać za wadę tego estymatora. Czasami wiadomo bowiem, że estymowana gęstość jest regularna, np. jest klasy $C^2(\mathbb{R})$. Tej wady pozbawiony jest estymator jądrowy gęstości.

2 Definicja i interpretacja

Niech x_1, x_2, …, x_n będzie próbą losową z rozkładu zmiennej losowej o nieznanej gęstości f. Estymator jądrowy gęstości, wyznaczony na podstawie tej próby, definiujemy wzorem:

$$\widehat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right), \tag{1}$$

gdzie $n \in \mathbb{N}$ oznacza liczność próby, ciąg dodatnich liczb (h_n) określony jest jako współczynnik wygładzania, który spełnia warunki $\lim\limits_{n\to\infty} h_n = 0$ oraz $\lim\limits_{n\to\infty} nh_n = \infty$, a $K\colon \mathbb{R} \to [0, \infty)$ jest mierzalną funkcją nazywaną jądrem, spełniającą następujące warunki:

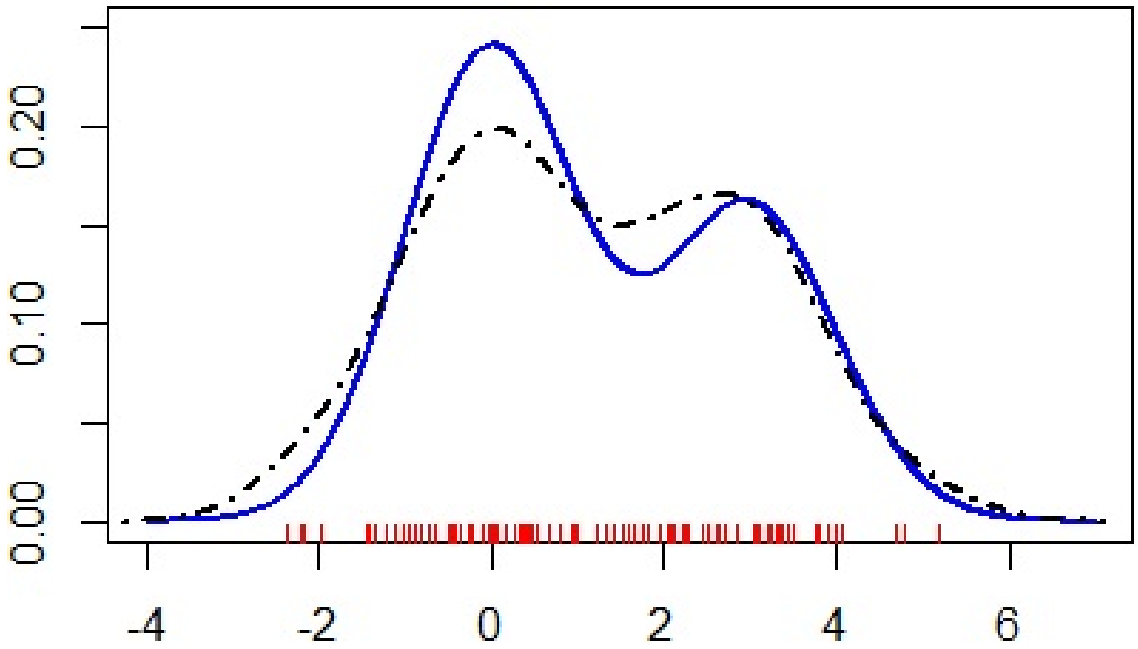
- $\int_{\mathbb{R}} K(x)\,dx = 1$,
- K jest symetryczna względem zera, tzn. $K(x) = K(-x) \quad \forall x \in \mathbb{R}$,
- K posiada w $x = 0$ słabe maksimum globalne, tzn. $K(0) \geqslant K(x) \quad \forall x \in \mathbb{R}$.

Te własności zapewniają, między innymi, że dla każdej realizacji próby, estymator jądrowy jest gęstością pewnego rozkładu prawdopodobieństwa.

W przypadku gdy mamy do czynienia z pojedynczą obserwacją x_i, funkcja K, przesunięta o wektor x_i oraz wygładzona parametrem h_n, przedstawia szacowany rozkład zmiennej losowej X, przy założeniu, że przyjęła ona wartość x_i. Kiedy mamy do czynienia z n–elementową realizacją x_1, x_2, …, x_n naszej zmiennej losowej X, rozkład wtedy jest szacowany przez sumę tych pojedynczych. Współczynnik $\dfrac{1}{nh_n}$ standaryzuje otrzymaną funkcję, aby zachodził warunek

$$\int_{\mathbb{R}} \widehat{f}(x)\,dx = 1, \tag{2}$$

co jest podstawowym żądaniem gęstości rozkładu prawdopodobieństwa.



Rysunek 2: Przykład jądrowego estymatora gęstości dla zmiennej losowej jednowymiarowej z użyciem jądra normalnego.

Na Rysunku 2 punkty, które oznaczają realizację próby pokazano kolorem czerwonym. Estymowaną gęstość przedstawiono czarnym przerywanym wykresem. Dla porównania, prawdziwą gęstość zmiennej losowej skąd pochodzi próba, zilustrowano ciemnoniebieskim ciągłym wykresem.

3 Wybór postaci jądra

Najpopularniejsze jądra jednowymiarowe razem ze wzorami.
Pierwszym z jąder, które zasługuje na uwagę jest tak zwane jądro Epanecznikowa. Wyraża się ono następującym wzorem:

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2), & x \in [-1,1], \\ 0, & x \in (-\infty,-1) \cup (1,\infty). \end{cases} \tag{3}$$

Jądro Epanecznikowa wywodzi się z pewnej rodziny jąder, które mają następującą postać:

$$K(x) = \begin{cases} \frac{1}{v_k}(1-x^2)^k, & x \in [-1,1], \\ 0, & x \in (-\infty,-1) \cup (1,\infty), \end{cases} \tag{4}$$

gdzie k jest liczbą całkowitą nieujemną, a v_k jest pewną stałą, którą można wyliczyć z następującego wzoru:

$$v_k = 2\int_0^1 (1-x^2)^k \, dx. \tag{5}$$

Zauważmy, że podstawiając $k = 0$ otrzymujemy tak zwane jądro jednostajne:

$$K(x) = \begin{cases} \frac{1}{2}, & x \in [-1,1], \\ 0, & x \in (-\infty,-1) \cup (1,\infty). \end{cases} \tag{6}$$

Kiedy podstawimy $k = 1$ otrzymamy wspomniane już jądro Epanecznikowa, natomiast gdy $k = 2$, dostajemy tak zwane jądro dwuwagowe:

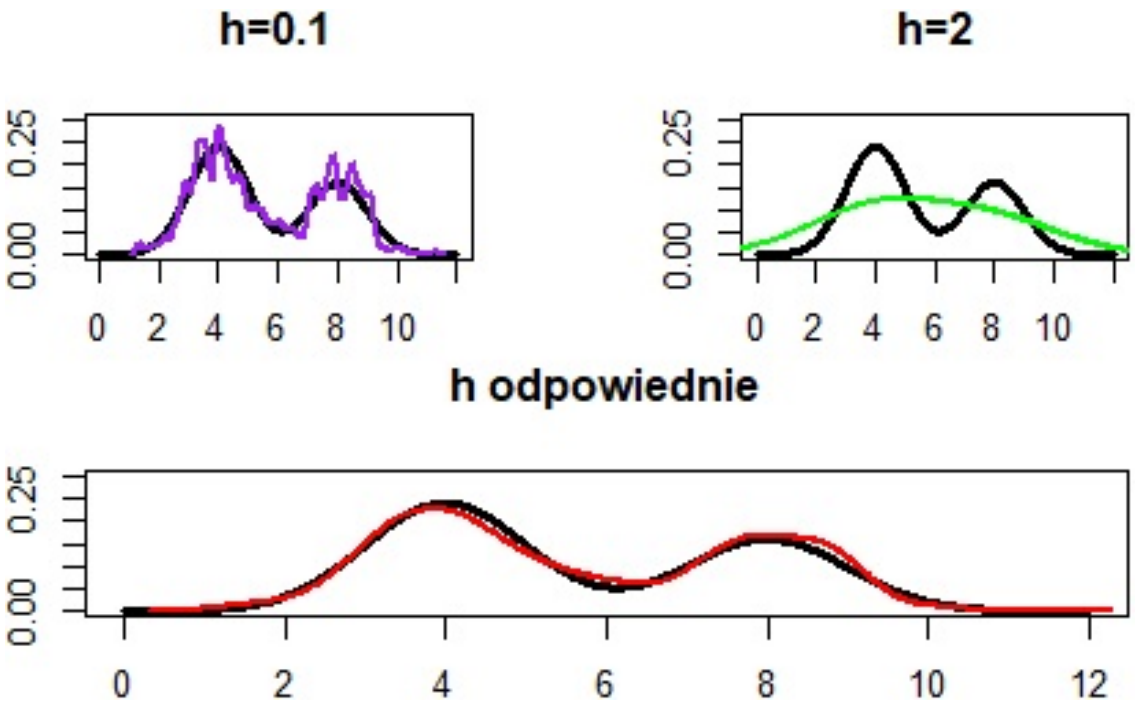
$$K(x) = \begin{cases} \frac{15}{16}(1-x^2)^2, & x \in [-1,1], \\ 0, & x \in (-\infty,-1) \cup (1,\infty). \end{cases} \tag{7}$$

Może dojść do sytuacji, w której zależy nam, aby nasz estymator $\widehat{f}$ posiadał pochodne dowolnego rzędu. W takiej sytuacji, najlepszy wybór to jądro normalne, czyli:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{8}$$

Czasami może się zdarzyć, iż posiadamy dodatkową informację skąd pochodzi nasza próba. Na przykład jeśli wiemy, że pochodzi ona z rozkładu normalnego, wtedy warto przyjąć jądro normalne do tworzenia naszego estymatora. W praktyce najczęściej wybierane jest właśnie jądro normalne, ponieważ funkcja e^x ma wiele dobrych właściwości.

4 Dobór parametru wygładzania



Rysunek 3: Lewy górny róg — zbyt mała wartość h. Prawy górny róg — zbyt duża wartość h. Na dole — odpowiednia wartość h. Kolorem czarnym na wszystkich trzech wykresach zaznaczono prawdziwą gęstość.

Ogólnie, standardowym kryterium oceny jakości estymatorów jest tzw. kryterium błędu średniokwadratowego. W przypadku gdy estymujemy parametr rzeczywisty b, gdzie naszym estymatorem jest $\widehat{b}$, wartość tego błędu, oznaczanego jako MSE, możemy policzyć ze wzoru:

$$\mathrm{MSE} = E((\widehat{b} - b)^2). \tag{9}$$

Wzór (9) możemy zapisać jako:

$$\mathrm{MSE} = \Big(E(\widehat{b}) - b\Big)^2 + \mathrm{Var}(\widehat{b}) = \Big[\mathrm{Bias}(\widehat{b})\Big]^2 + \mathrm{Var}(\widehat{b}). \tag{10}$$

Wartość MSE jest więc sumą kwadratu obciążenia estymatora $\widehat{b}$ oraz jego wariancji. W przypadku gdy estymujemy gęstość rozkładu prawdopodobieństwa, wartość MSE w ustalonym punkcie $x \in \mathbb{R}$ możemy obliczyć następująco:

$$\mathrm{MSE_x} = E\left([\widehat{f}(x) - f(x)]^2\right) \tag{11}$$

lub równoważnie:

$$\mathrm{MSE_x} = \Big(E(\widehat{f}(x)) - f(x)\Big)^2 + \mathrm{Var}(\widehat{f}(x)) = \Big[\mathrm{Bias}(\widehat{f}(x))\Big]^2 + \mathrm{Var}(\widehat{f}(x)), \tag{12}$$

gdzie $\widehat{f}$ jest estymatorem gęstości rozkładu naszej zmiennej losowej, o gęstości prawdopodobieństwa f. Całkowitą jakość estymacji możemy zmierzyć całkując $\mathrm{MSE_x}$, po całej przestrzeni zmiennej losowej. Wartość takiej operacji, określana jest jako scałkowany błąd średniokwadratowy i jest równa:

$$\mathrm{MISE} = \int E\left([\widehat{f}(x) - f(x)]^2\right) dx. \tag{13}$$

Dzięki temu możemy otrzymać teoretycznie optymalną wartość parametru wygładzania, tzn.

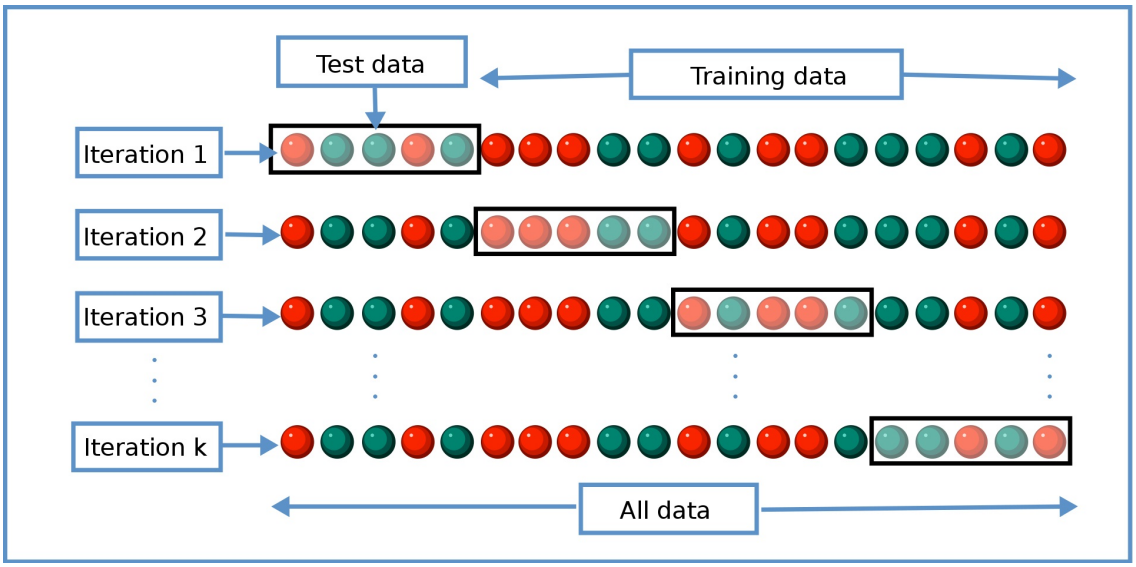
$$h_{opt} = \left(\frac{W(K)}{(U(K))^2 Z(f)} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}. \tag{14}$$

gdzie $U(K) = \int x^2 K(x)\, dx, \quad W(K) = \int K^2(x)\, dx, \quad Z(f) = \int (f''(x))^2\, dx.$

Niestety jednak, ta wartość zależy od nieznanej wartości $Z(f) = \int (f''(x))^2\, dx$, dlatego w większości sytuacji powyższy wzór nie jest przydatny. Możemy go jednak wykorzystać, jeśli wiemy, z jakiej rodziny rozkładów pochodzi nasza próba. Wtedy możemy estymować nieznaną wartość $Z(f)$. Skutkiem tego, otrzymamy optymalną wartość parametru wygładzania.

5 Cross–Validation

Kolejną dosyć znaną metodą znalezienia optymalnego h jest cross–validation. Jej pomysł opiera się na dzieleniu naszej próby na $k \in \mathbb{N}$ równolicznych (ewentualnie niemal równolicznych) rozłącznych pozdbiorów. Po tym podziale tworzy się 2 rozłączne zbiory; zbiór uczący i testowy. Zbiór uczący składa się z $k-1$ utworzonych wcześniej podzbiorów, a zbiorem testowym jest jeden nie wybrany. Na zbiorze uczącym tworzy się model z ustaloną pewną szerokością okna, następnie sprawdza się jak ten model pasuje do zbioru testowego, licząc wartość błędu MISE. Procedurę tworzenia zbiorów uczącego i testowego powtarza się k razy, za każdym razem zmieniając zbiór testowy. Każdej badanej szerokości okna przypisuje się wartość zsumowanych błędów MISE uzyskanych na k zbiorach testowych. Ostatecznie trzeba wybrać to h, dla którego wartość sum MISE jest najmniejsza.



Rysunek 4: Jak rozumieć cross–validation