معرفی روش‌های اصلاح کران در برآورد ناپارامتری تابع چگالی و رگرسیون هسته

الهه آخوندزاده کاشانی، مهدیه میرزایی باغینی
گروه آمار، دانشگاه فردوسی مشهد

چکیده

در این مقاله پس از معرفی برآوردگرهای توابع چگالی و رگرسیونی به روش ناپارامتری هسته‌ای، اثرات کران معرفی خواهد شد و در ادامه به معرفی روش‌های اصلاح اثرات کران می‌پردازیم. در انتها نیز دو دلیل استفاده‌ی محدود از این روش‌ها را معرفی خواهیم کرد.

واژه‌های کلیدی: اثرات کران، برآورد ناپارامتری رگرسیون، روش هسته‌ای.

۱ مقدمه

در برآورد توابع چگالی و رگرسیون به روش ناپارامتری افزایش واریانس و اریبی در کران‌ها، که اصطلاحاً اثرات کران نامیده می‌شوند، می‌تواند کاملاً مشکل‌ساز باشد و حتی ممکن است برآوردهای نهایی و در نهایت نتایج را اساساً تغییر دهد. در بسیاری از مطالعات کاربردی کران‌ها مورد توجه خاص هستند، برای مثال در تحلیل فقر، لازم است که برآوردهای قابل اعتمادی از توزیع درآمد در سمت چپ یعنی نزدیک به کران طبیعی صفر داشته باشیم. به طور مشابه وقتی از رگرسیون ناپارامتری در اقتصاد سنجی استفاده می‌شود، اثرات کران نه تنها در کران‌ها بلکه در نزدیک آن‌ها هم دیده می‌شوند. با توجه به این مثال‌ها می‌توان مشاهده کرد که استفاده از روش‌های اصلاح کران در هر دو مورد برآورد هسته‌ای توابع چگالی و رگرسیونی ضرورت دارد.

تاکنون روش‌های متعددی برای اصلاح اثرات کران معرفی شده‌اند. در این مقاله چند روش اصلاح اثرات کران معرفی شده‌اند که در ادامه به معرفی چند مورد از آن‌ها می‌پردازیم. ابتدا، برای فهم بهتر مطالب مربوط به اثرات کران، به یادآوری مختصری از بحث برآورد ناپارامتری توابع چگالی و رگرسیونی به روش هسته‌ای خواهیم پرداخت و در ادامه پس از معرفی کلی اثرات کران، به بحث و بررسی بیشتری درباره‌ی روش‌های اصلاح اثرات کران پرداخته می‌شود.

۲ برآوردگر هسته‌ای تابع چگالی

فرض کنید نمونه تصادفی $X_1, X_2, \ldots, X_n$ را به حجم n از توزیعی پیوسته و نامعلوم داریم. در این صورت برآورد تابع چگالی f در فاصله‌ی $[x-h, x+h]$ به طول $2h$ توسط $\hat{f}_h(x)$ در عبارت (۱) محاسبه می‌شود.

$$\hat{f}_h(x) = \frac{1}{2nh} \# \{X_i \in [x-h, x+h]\}, \qquad (1)$$

که در آن # به مفهوم تعداد مشاهدات قرار گرفته در بازه مورد نظر می‌باشد.
اگر در این عبارت از تابع وزنی $K(u) = \frac{1}{2} I(|u| \le 1)$ ، $(u = (x - X_i)/h)$ استفاده شود، می‌توان عبارت (۱) را به صورت عبارت (۲) بازنویسی کرد:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \qquad (2)$$

$$= \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} I\left(\left|\frac{x - X_i}{h}\right| \le 1\right). \qquad (3)$$

K مختصرنویسی یک تابع وزنی به نام تابع هسته‌ی یکنواخت است که به هر مشاهده X_i که فاصله آن از x (نقطه‌ای که می‌خواهیم تابع چگالی احتمال را برآورد کنیم) کمتر از h است وزن ۱/۲ را تخصیص می‌دهد و مشاهداتی که فاصله آن‌ها از x بیشتر از h است وزن صفر می‌گیرند.

توابع هسته‌ی متقارن توابعی با تکیه‌گاه $[-1, 1]$ هستند که در $\mu_0(K) = 1$ و $\mu_1(K) = 0$ با $\mu_l(K) = \int_{-1}^{1} u^l K(u) du (l = 0, 1, 2; u = \frac{X-x}{h})$ صدق می‌کنند. تاکنون توابع هسته‌ی متعددی معرفی شده‌اند که تعدادی از آن‌ها در جدول ۱ معرفی شده‌اند.

جدول ۱: توابع هسته مختلف

هسته	$\mathbf{K(u)}$
یکنواخت	$\frac{1}{2} I(\lvert u\rvert \le 1)$
مثلث	$(1 - \lvert u\rvert) I(\lvert u\rvert \le 1)$
اپنیچ نیکوف	$\frac{3}{4}(1 - u^2) I(\lvert u\rvert \le 1)$
درجه‌ی چهارم (دو وزنی)	$\frac{15}{16}(1 - u^2)^2 I(\lvert u\rvert \le 1)$
سه وزنی	$\frac{35}{32}(1 - u^2)^3 I(\lvert u\rvert \le 1)$
گاوسی	$\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}u^2\right)$
کسینوس	$\frac{\pi}{4}\cos(\frac{\pi}{2}u) I(\lvert u\rvert \le 1)$

اکنون می‌توان فرم کلّی برآوردگر هسته‌ای تابع چگالی احتمال در نقطه x را معرفی نمود. بر همین اساس با توجه به یک نمونه تصادفی $X_1\ ,\ X_2\ ,\ \ldots,\ X_n$ از f داریم:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \qquad (4)$$

که

$$K_h(\cdot) = \frac{1}{h} K(\cdot/h), \qquad (5)$$

$K(\bullet)$ یک تابع هسته مشابه آنچه در جدول ۱ نمایش داده شده است، می‌باشد و h پهنای باند را مشخص می‌کند. توجه کنید که عبارت "تابع هسته" اشاره به تابع وزنی K دارد، در حالی که برآوردگر هسته‌ای چگالی اشاره به فرمول (۴) دارد.

۳ رگرسیون هسته‌ای

سوال مهمّی که در بسیاری از زمینه‌های علمی مطرح است، چگونگی رابطه‌ی بین دو متغیّر می‌باشد که X و Y نامیده می‌شوند. تحلیل رگرسیونی به این سوال که چگونه Y (متغیّر وابسته) می‌تواند توسط X (متغیّر مستقل یا

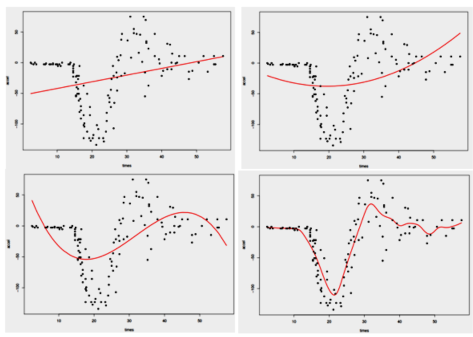
توضیحی) توضیح داده شود، پاسخ می‌دهد. این به معنی رابطه‌ای به فرم :

$$Y = m(X),$$

است که $m(\bullet)$ در مفهوم ریاضی یک تابع است. اکثراً در فرضیّات علمی هیچ قیدی بر فرم تابعی $m(\bullet)$ تحمیل نمی‌شود. یعنی، در فرضیّات بیان نمی‌شود که آیا $m(\bullet)$ خطّی، درجه دو، نسبت به X صعودی یا غیره است. از این‌رو یافتن اطّلاعات بیشتر در مورد $m(\bullet)$ تا حدودی یک تحلیل تجربی است.

خاطر نشان می‌کنیم که در روش‌های پارامتری اغلب فرض می‌شود $m(x) = \alpha + \beta x$ و مسئله برآورد $m(x)$ به مسئله برآورد α و β ساده می‌شود. امّا این روش در برخی موارد که داده‌ها فرم تابعی پیچیده‌ای دارند مفید نیست، از این‌رو در روش‌های ناپارامتری چنین قیدهایی روی $m(\bullet)$ در نظر گرفته نمی‌شود.

مثال ۱.۳. *به منظور درک بهتر این مسئله شکل ۱ را در نظر بگیرید. این نمودار مربوط به داده‌های تصادفات موتورسواران در نرم‌افزار R می‌باشد.*



شکل ۱: برازش خطی(بالا سمت چپ)، برازش درجه دو (بالا سمت راست) رگرسیون درجه سه (پایین سمت چپ)، رگرسیون هسته‌ای (پایین سمت راست)

نتایج شکل ۱ مؤید این نکته است که رگرسیون ناپارامتری بدون در نظرگرفتن هیچ‌گونه قیدی روی $m(\bullet)$ بهترین نمودار را به داده‌ها برازش می‌دهد.

هنگامی‌که X و Y هر دو متغیرهای تصادفی با تابع چگالی احتمال توأم $f(x,y)$ باشند، طرح نمونه‌گیری برای استخراج n جفت $\{(X_i,Y_i)\}$ ، طرح تصادفی نام دارد، از این‌رو می‌توان با تعریف امید شرطی، برآوردگر رگرسیونی رابه صورت زیر محاسبه کرد:

$$\begin{aligned} m(x) &= E(Y \mid X = x) \\ &= \int y \frac{f(x,y)}{f_X(x)} dy = \frac{\int y f(x,y)\, dy}{f_X(x)}, \end{aligned} \qquad (۶)$$

با این فرض که مشاهداتی به فرم $\{X_i,Y_i\}, i = 1,2,\dots,n$ داریم، تنها کمّیت‌های نامعلوم در سمت راست عبارت (۶)، $f(x,y)$ و $f_X(x)$ هستند. با توجه به مباحث مربوط به برآورد چگالی هسته، می‌دانیم چگونه توابع چگالی احتمال را برآورد کنیم. برآورد $f_X(x)$ سرراست است (با استفاده از عبارت (۴)). برای برآورد $f(x,y)$ می‌توان از برآوردگر چگالی چند-متغیّره با هسته ضربی که به صورت زیر تعریف می‌شود، استفاده کرد.

$$\kappa(u) = K(u_1)\dots\ K(u_d),$$

که در آن $\kappa(u) = \kappa(u_1, \dots, u_d)$ تابع هسته‌ی چند بعدی است و $K(u_1),\dots,K(u_d)$ معرف توابع هسته یک متغیّره هستند. از این‌رو داریم:

$$\begin{aligned} &\hat{f}_{h,g}(x,y) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_h\left(\frac{x-X_i}{h}\right) K_g\left(\frac{y-Y_i}{g}\right). \end{aligned} \qquad (۷)$$

بر اساس برآورد معرفی شده برای تابع چگالی توأم به

روش هسته، می‌توان نشان داد:

$$\int y\hat{f}_{h,g}\left(x,y\right)dy=\frac{1}{n}\sum_{i=1}^{n}K_h\left(x-X_i\right)Y_i$$

با قرار دادن عبارت بالا در معادله (۶)، برآوردگر به دست می‌آید که توسط نادار‌یا [۶] و واتسون [۱۰] به‌صورت زیر معرفی شد:

$$\hat{m}_h\left(x\right)=\frac{n^{-1}\sum_{i=1}^{n}K_h(x-X_i)Y_i}{n^{-1}\sum_{j=1}^{n}K_h(x-X_j)}. \qquad (8)$$

۴ روش‌های اصلاح کران

در این بخش به معرفی برخی از مهمترین روش های اصلاح کران پرداخته می شود.

۱.۴ اثرات کران

فرض کنید می‌خواهیم تابع چگالی احتمال f را به صورت ناپارامتری براساس یک نمونه تصادفی به حجم n یعنی به‌صورت $x_1,x_2,\ldots,x_n$ برآورد کنیم. با توجه به مطالب ذکر شده در بخش ۲، برآوردگر هسته‌ای چگالی استاندارد f توسط معادله‌ی (۴) محاسبه می‌شود. برای قابل قبول بودن روش هسته فرض می‌شود f تابعی هموار است و مشتق مرتبه‌ی دوم آن موجود است. اریبی و واریانس برآوردگر چگالی هسته و برآوردگر نادار‌یا و واتسون برای تابع رگرسیونی طبق قضایای زیر محاسبه می‌شود.

قضیه ۱.۴. *اریبی و واریانس برآوردگر هسته‌ای چگالی به صورت زیر محاسبه می‌شود (اثبات در [۱]):*

$$Bias\left\{\hat{f}_h\left(x\right)\right\}=\frac{h^2}{2}f''\left(x\right)\mu_2\left(K\right)+o\left(h^2\right),\ as\ \ h\to 0 \qquad (9)$$

$$Var\left\{\hat{f}_h\left(x\right)\right\}=\frac{1}{nh}\|K\|_2^2f\left(x\right)+o\left(\frac{1}{nh}\right),\quad as\quad nh\to\infty \qquad (10)$$

که در‌آن $\mu_2\left(K\right)=\int s^2K(s)ds$ و $\|K\|_2^2=\int K^2(s)ds$.

قضیه ۲.۴. *اریبی و واریانس برآوردگر نادار‌یا-واتسون به صورت زیر محاسبه می‌شود (اثبات در [۱]):*

$$Bias\left(\hat{m}\right)=\frac{h^2\mu_2\left(K\right)}{2f\left(x\right)}(m''(x)f(x)+f'(x)m'(x))+O\left(\frac{1}{nh}\right)+o(h^2)$$

$$V\left(\hat{m}\right)=\frac{\sigma^2(x)}{nhf(x)}\|K\|_2^2+O\left(\frac{1}{nh}\right),$$

که $\sigma^2\left(x\right)=V[Y_i|X_i=x]$.

در این‌جا ذکر این نکته لازم است که در نواحی کرانی اریبی از مرتبه‌ی $o(h)$ است که تا حدّی بدتر از اریبی مرتبه‌ی $o(h^2)$ در ناحیه درونی تکیه‌گاه است.
در تمام این محاسبات (برآورد چگالی و برآورد رگرسیونی) فرض شده است که تکیه‌گاه تابع چگالی دامنه اعداد حقیقی باشد. اگر این فرض برقرار نباشد و به عنوان مثال برآورد هسته‌ای تابع چگالی برای داده‌ها روی اعداد حقیقی مثبت انجام گیرد، نزدیک کران تکیه‌گاه (در مبدأ مختصات) واریانس و اریبی برآوردگر افزایش می‌یابد. این مسئله بدین علّت است که برآوردگر هسته‌ای تابع چگالی دانشی در مورد کران ندارد و به طور کلّی، جرم احتمال را به خارج از تکیه‌گاه اختصاص می‌دهد و برای تمام نقاط x که در کران‌ها قرار دارند، جستجوی هسته سهواً برای

اطلاعات خارج از تکیه‌گاه f صورت می‌گیرد. افزایش واریانس و اریبی در کران‌ها در اصطلاح "اثرات کران" نامیده می‌شود. در مورد رگرسیون ناپارامتری نیز، مسئله اثرات کران مشهود است و در عمل می‌تواند به‌عنوان یک مسئله جدی مطرح شود و برآورد صورت گرفته را تحت تاثیر قرار دهد.
به‌عنوان مثال در مورد بررسی فقر در یک جامعه، لازم است تا برآورد قابل قبولی از توزیع درآمد در سمت چپ، نزدیک کران طبیعی صفر، داشته باشیم. به همین ترتیب اگر علاقه‌مند به تحلیل ریسک، فقر و نابرابری باشیم به ترتیب باید به گروه های سنی (جوان یا پیر)، تفاوت در میزان تحصیلات و مقایسه‌ی شرکت‌های بزرگ و کوچک و ... بپردازیم. بنابراین در موارد بسیاری روی کران‌ها تمرکز داریم.
در این مقاله نواحی کرانی چپ و راست به ترتیب توسط B_l و B_r نمایش داده می‌شوند و به صورت زیر تعریف می‌شوند:

$$B_l = \{\mathrm{x} : \mathrm{a} \le \mathrm{x} < (a+h)\},$$
$$B_r = \{\mathrm{x} : (c-h) < x \le c\},$$

که در آن (h) به پهنای باند اشاره دارد و $a-$ در صورت وجود، کران پایین است و کران بالا، در صورت وجود توسط c نمایش داده می‌شود، از این‌رو $[a+h \ , \ c-h]$ ناحیه درونی می‌باشد.

تاکنون روش‌های متعددی برای اصلاح اثرات کران معرفی شده‌اند، به طور کلی روش‌های موجود می‌تواند به گروه‌های زیر تقسیم شود:

- روش تبدیل داده‌ها (واند و همکاران[1][۹])
- روش انعکاس داده‌ها (اسکاتر[2][۸])
- روش شبه داده (کولینگ و هال[3][۲])
- روش هسته‌های کرانی (ژانگ و کارونامونی[4][۱۱])
- روش پهنای باند محلی (رایس[5][۷])

۲.۴ روش تبدیل داده‌ها

در این روش برآوردگر هسته‌ای چگالی به صورت زیر تعریف می‌شود:

$$\hat{f}_h(x) = \frac{۱}{nh}\sum_{i=۱}^{n} K\left(\frac{x-g(X_i)}{h}\right),$$

که در آن $g : [۰ \ , \ \infty) \to [۰ \ , \ \infty)$ تابعی پیوسته و یک به یک است و K برآوردگر هسته‌ای معمولی می‌باشد. توجه کنید که این برآوردگر در حقیقت برآوردگر تابع چگالی احتمال X نیست، بلکه برآوردگر $g(X)$ است.

۳.۴ روش انعکاس داده‌ها

با توجه به مطالب ذکر شده در بخش ۱.۴، اگر برآورد هسته ای تابع چگالی برای داده ها روی اعداد حقیقی مثبت انجام گیرد، به دلیل عدم وجود داده در محورمنفی اثرات کران افزایش می یابد. از این رو، ایده اصلی این روش براساس اضافه کردن $-X_۱, -X_۲, \ldots, -X_n$ به مجموعه‌ی داده‌هاست، بنابراین می‌توان تابع چگالی را در این روش با استفاده از عبارت (۱۱) برآورد کرد.

$$\hat{f}(x) = \begin{cases} \frac{۱}{nh}\sum_{i=۱}^{n}\left\{K\left(\frac{x-X_i}{h}\right)\right. & \\ \left.+K\left(\frac{x+X_i}{h}\right)\right\} & x \ge ۰ \\ ۰ & x < ۰ \end{cases} \qquad (۱۱)$$

[1] Wand et al.
[2] Schuster
[3] Cowling and Hall
[4] Zhang and Karunamuni
[5] Rice

۴.۴ روش شبه داده

در این زیربخش، روش شبه داده کولینگ و هال [۲] معرفی می‌شود که در واقع نوعی از برآوردگر تبدیل یافته‌ی انعکاسی است، در این روش ابتدا داده‌ها به مجموعه‌ای جدید تبدیل می‌شوند و سپس این داده‌ها در محور منفی قرار می‌گیرند. کولینگ و هال برآوردگر تابع چگالی را به‌صورت زیر معرفی کردند:

$$\hat{f}(x) = \frac{1}{nh}\left\{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) + \sum_{i=1}^{m} K\left(\frac{x-X_{-i}}{h}\right)\right\}, \qquad (12)$$

که m به گونه‌ای است که $O(nh) < O(m) < O(n)$ و همچنین متغیر X_{-i} برای ثابت‌هاي مثبت $A_1, ..., A_s (s \geq r)$ که r به همواری تابع چندك X در لبه‌ی مورد نظر وابسته است) و اعداد حقیقي $a_1, ..., a_s$، به‌صورت زیر تعریف می‌شود:

$$X_{-i} = \sum_{j=1}^{s} a_j X_{A_j i}\ ,\ 1 \leq i \leq \frac{n}{\max\{A_i\}}$$

به طوری‌که

$$\sum_{k=1}^{s} a_k A_k^j = (-1)^j, \quad 1 \leq j \leq r \qquad (13)$$

در مقاله کولینگ و هال [۲] X_{-i} توسط قانون "بهترین سه نقطه" به صورت زیر محاسبه می‌شود:

$$X_{-i} = -5X_{(i/3)} - 4X_{(2i/3)} + 10/3 X_{(i)} \quad , i = 1\ ,\ 2\ ,\ \ldots,\ n.$$

۵.۴ روش هسته‌های کرانی

ژانگ و کاروناموني [۱۱] هسته‌های کرانی با پهنای باند متغیر را به صورت زیر معرّفی کردند. در این روش برای برآورد تابع چگالی در هر نقطه‌ای که در ناحیه کرانی قرار بگیرد از یک هسته‌ی متفاوت استفاده می‌شود که معمولاً هسته‌های جدید فاقد ویژگی تقارن هستند و روی محور مثبت وزن بیشتری می‌گیرند. بر این اساس برآوردگر به صورت زیر نمایش داده می‌شود:

$$\hat{f}(x) = \frac{1}{nh_c}\sum_{i=1}^{n} K_{(c/b(c))}\left(\frac{x-X_i}{h_c}\right), \qquad (14)$$

که در رابطه بالا از $c = \min\{x/h, 1\}$ در ناحیه کرانی استفاده می‌شود و $b(c) = 2 - c$ است. همچنین هسته‌ی کرانی استفاده شده در رابطه فوق عبارت است از:

$$K_{(c)}(t) = \frac{12}{(1+c)^4}(1+t)\left\{(1-2c)\,t + \frac{3c^2-2c+1}{2}\right\} I\{-1 \leq t \geq \mathbf{c}\}.$$

ژانگ و کاروناموني [۱۱] نشان دادند که این هسته در زمینه‌ی مینیمم ساختن MSE بین کلاس تمامی هسته‌ها بهینه است.

۶.۴ روش پهنای باند محلی

یک روش دیگر، انتخاب پهناهای باند محلی در ناحیه کران است. به ویژه ممکن است این نظر وجود داشته باشد که در نواحی کرانی باید از پهناهای باند بزرگ‌تر استفاده کرد. رایس [۷] و گسر و همکاران [۳] پهنای باندی را پیشنهاد دادند که در کران ثابت است. برای دستیابی به این مسئله، در حقیقت برای تمام نقاط کران

یک پهنای باند محلی استفاده می‌شود که به‌صورت زیر تعریف می‌شود:

$$h_x = \begin{cases} ۲h_{global} - (x-a) & a < x < (a + h_{global}), \\ ۲h_{global} - (c-x) & (c - h_{global}) < x < c, \\ h_{global} & \text{سایر جاها} \end{cases}$$

(۱۵)

هال و ورلی [۴] این ایده را ابتدا بر روی داده‌های شبیه‌سازی شده توسعه دادند (با یک نوع از بوت-استرپ) و سپس برآورد را در ناحیه کران با استفاده از مجموعه‌ای از داده‌های حقیقی و داده‌های فرضی قبلی انجام دادند.

در زمینه‌ی برآورد تابع رگرسیونی $m(\bullet)$ ، رایس (۱۹۸۴) برای تمام نقاط کرانی $x = a + h\rho$، $\rho < ۱$ برآوردگر (۱۶) را پیشنهاد کرد که ترکیبی خطی از برآورگرهای اصلاح نشده‌ی $\hat{m}_{h_{global}}$ و برآوردگرهای اصلاح شده‌ی $\hat{m}_{h_x}$ است.

$$\tilde{m}(x) = (۱ + \beta_\rho)\,\hat{m}_{h_{global}}(x) - \beta_\rho \hat{m}_{h_x}(x), \quad (۱۶)$$

که $\hat{m}$ در (۸) تعریف شده و h_x بر اساس رابطه (۱۵) انتخاب می‌شود.

$$\beta_\rho = \frac{w_۱(\rho)\, w_۰^{-۱}(\rho)}{(۲-\rho)\, w_۱\left(\frac{\rho}{۲-\rho}\right) - w_۱ w_۰^{-۱}},$$

$$w_k(v) = \int_{-۱}^{v} u^k K(u)\, du .$$

۷.۴ روش جونز

همانگونه که اشاره شد اکنون روش‌های اصلاح کران متعددی برای برآوردگر هسته‌ای چگالی موجود است اما اغلب این روش‌ها با روش اصلاح کرانی که توسط جونز[۶] معرّفی شد، مقایسه می‌شوند. در این بخش روش جونز معرفی شده و ویژگی‌های واریانس و اریبی آن نیز نمایش داده می‌شود. جونز پیشنهاد کرد که هسته‌ی استفاده شده در این برآورد (به ویژه اگر تابع f روی $[a,c]$ تعریف شده باشد) به‌صورت زیر تعریف می‌گردد:

$$K^*(u) = \frac{w_۳ - w_۲ u}{w_۱ w_۳ - (w_۲)^۲} K(u)\, 1_{[c_۲, c_۱]}, \quad (۱۷)$$

به عبارت دیگر جونز پیشنهاد کرد که توان بیشتری از درون تکیه‌گاه بگیریم، به طوری‌که مؤلفه‌های دوباره نرمال شده‌ی w_j به‌صورت زیر تعریف می‌گردد:

$$w_j = \int_{c_۲}^{c_۱} \left(\frac{t-x}{h_{global}}\right)^{j-۱} K\left(\frac{t-x}{h_{global}}\right) dt ,$$

که $c_۱ = min(c, x + h_{global})$ و $c_۲ = \max(a, x - h_{global})$ می‌باشند. پس برآورد تابع چگالی بر اساس این روش جهت اصلاح اثرات کران، با به کارگیری هسته به طور خطی اصلاح شده $K^*(u)$ در رابطه (۴) به دست می‌آید. به طور مشابه برای برآورد رگرسیونی (۱۱) از $K^*(u)$ در تعریف $\hat{m}_h(x)$ استفاده خواهیم کرد.

حال مجانب‌های برآوردگر هسته وقتی روش جونز [۵] به کار رفته است را در نظر بگیرید. بدون کاستن از کلیّت فرض کنید کران پایین a را داریم، به خاطر دارید که پذیرفتیم هسته‌ها روی $[-۱, ۱]$ کراندار هستند. در این‌جا اندیس $global$ از پهنای باند حذف شده است و ضمناً اسکالر p را که به x و a از طریق $x = p(a+h)$ وابسته

[۶] Jones

است، تعریف شده است. پس برای

$$a_l(p) = \int_{-1}^{\min[1,p]} u^l K(u)du$$

و

$$b(p) = \int_{-1}^{\min[1,p]} K^2(u)du,$$

مجانب‌ها می‌توانند توسط معادلات زیر تخمین زده شوند:

$$Bias\left\{\hat{f}_h(x)\right\} \simeq f(x)\left(a_0(p) - 1\right) - ha_1(p) f'(x) + \frac{h^2}{2} f''(x) a_2(p) \quad , \tag{۱۸}$$

$$var\left\{\hat{f}_h(x)\right\} \simeq \frac{1}{nh} f(x) b(p). \tag{۱۹}$$

توجه داشته باشید که برای تمام نقاط درونی، مجانب‌ها به ترتیب با عبارات عمومی (۹) و (۱۰) برابرند.

در نهایت می‌توان ذکر کرد که علیرغم اهمیت زیاد روش‌های اصلاح کران در عمل و همچنین اهمیت قابل توجه (اما نه زیاد) آن‌ها در مطالعات نظری، این روش‌ها هم در برآورد چگالی و هم در رگرسیون به ندرت استفاده می‌شوند. یک دلیل واضح این امر، فقدان برنامه‌های اجرایی در نرم‌افزارهای آماری و اقتصادی است. در کنار فقدان نرم‌افزار دلیل دیگر برای استفاده‌ی اندک از این روش‌ها می‌تواند پیچیدگی آن‌ها در مقایسه با میزان بهبود مشهود ایجاد شده در برآورد نهایی باشد.

مراجع

[۱] آخوندزاده کاشانی، ا. (۱۳۹۱). معرفی روشی ساده و موثر برای اصلاح کران در برآورد ناپارامتری تابع چگالی و رگرسیون هسته. پایان نامه کارشناسی ارشد، دانشگاه فردوسی مشهد.

[2] Cowling, A. and Hall, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation. Journal of the Royal Statistical Society, Series B ,58 (3), 551-563.

[3] Gasser, T., Müller, H. G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. Journal of the Royal Statistical Society, Series B 47 (2), 238-252.

[4] Hall, P. and Wehrly, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. Journal of the American Statistical Association, 86 (415), 665-672.

[5] Jones, M. C. (1993). Simple boundary correction in kernel density estimation. Statistics and Computing, 3, 135-146.

[6] Nadaraya, E. A. (1964). On estimating regression. Theory of Probability and its Applications, 10, 186-190.

[7] Rice, J. (1984). Boundary modification for kernel regression.Communications in Statistics Theory and Methods,13 (7), 893-900.

[8] Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics. Theory and Methods, 14(5), 1123-1136.

[9] Wand, M. P., Marron, J. S. and Ruppert, D. (1991). Transformations in density estimation. Journal of the American Statistical Association, 86 (414), 343-353.

[10] Watson, G. S, (1964). Smooth Regression Analysis, Sankhya, 26:15, 175-184.

[11] Zhang, S. and Karunamuni, R. J. (1998). On kernel density estimation near endpoints. Journal of Statistical Planning and Inference, 70, 301-316.