مجله علوم آماری، بهار و تابستان ۱۳۹۰
جلد ۵، شماره ۱، ص ۴۱-۶۰

برآورد بیزی و اعتبارسنجی متقابل پهنای باند برآوردگر هسته‌ای تابع چگالی برای داده‌های در طول-اریب

مسعود عجمی بختیاروند [۱]، وحید فکور [۱]، سارا جمهوری [۲]
[۱]**گروه آمار، دانشگاه فردوسی مشهد**
[۲]**گروه آمار، دانشگاه بیرجند**

تاریخ دریافت: ۱۳۹۰/۲/۱۰ تاریخ آخرین بازنگری: ۱۳۹۰/۵/۱۶

چکیده : چنانچه در نمونه‌گیری، داده‌ها با احتمالی متناسب با اندازه انتخاب شوند، داده‌های حاصل را در طول-اریب نامند. برآورد ناپارامتری تابع چگالی با استفاده از داده‌های در طول-اریب، مشکل‌تر از سایر حالات است. یکی از برآوردگرهای معروف در این زمینه توسط جونز (۱۹۹۱) معرفی شده است. در این مقاله ابتدا پارامتر پهنای باند این برآوردگر با رهیافت بیزی برآورد می‌شود. سپس سازگاری قوی آن با به کار بردن پهنای باند برآورد شده به روش بیزی اثبات می‌شود. در انتها بامطالعه شبیه سازی به مقایسه عملکرد روش بیزی و اعتبارسنجی متقابل در برآورد پهنای باند پرداخته می‌شود.

واژه‌های کلیدی : برآوردگر هسته‌ای چگالی، پهنای باند، داده‌های در طول-اریب، اعتبارسنجی متقابل کمترین توان‌های دوم، هسته گاوسی وارون.

آدرس الکترونیک مسئول مقاله: مسعود عجمی بختیاروند، ajami.masoud@yahoo.com
کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲G۰۷

۱ مقدمه

پدیده در طول-اریب[۱]، اولین بار توسط ویکسل (۱۹۲۵) در آناتومی مطرح شد که وی نام آن را مسأله گلبولی نهاد. او در هنگام دیدن گلبول‌ها در میکروسکوپ متوجه شد که فقط گلبول‌هایی قابل مشاهده هستند که اندازه بزرگی آن‌ها، از حد معینی بیشتر باشد و گلبول‌های کوچکتر قابل دیدن در میکروسکوپ نیستند. بعدها این موضوع توسط مک فادن (۱۹۶۲) ، بلومنتال (۱۹۶۷) و کاکس (۱۹۶۹) به مفهوم آماری مورد بررسی قرار گرفت . کاکس (۱۹۶۹) در نمونه گیری یک نوع محصول صنعتی، متوجه شد که الیاف‌های با طول بلندتر، با احتمال بیشتری وارد نمونه می‌شوند. این موضوع نوعی اریبی را به نتایج تحمیل کرد که به "در طول- اریبی" معروف شد. در حالت کلی، اگر در نمونه‌گیری، عناصری از جامعه که اندازه، طول یا عمر بیشتری نسبت به بقیه اعضا جامعه دارند، وارد نمونه شوند یا شانس ورود به نمونه آن‌ها بیشتر باشد یا به عبارت دیگر عناصر جامعه با احتمالی متناسب با اندازه طول‌شان وارد نمونه شوند، نمونه دچار نوعی اریبی به نام در طول- اریبی می‌شود.

تعریف ۱ : فرض کنید F یک تابع توزیع تجمعی مطلقاً پیوسته با تابع چگالی f باشد. متغیر تصادفی Y را در طول-اریب گویند هرگاه تابع توزیع آن به صورت

$$G(y) = \int_0^y \frac{t}{\mu} dF(t), \quad y \geq 0, \tag{۱}$$

باشد، که در آن $\mu = \int_0^\infty t f(t) dt < \infty$.

با توجه به رابطه (۱) چگالی متغیر تصادفی Y به صورت زیر است

$$g(y) = \frac{y f(y)}{\mu}, \quad y \geq 0. \tag{۲}$$

در واقع رابطه (۲) بیانگر آن است که احتمال مشاهده هر مقدار متغیر تصادفی Y متناسب با طول آن است. معمولاً تابع چگالی f را چگالی نااریب و g را چگالی اریب گویند. در مسأله برآورد تابع چگالی f با داده‌های در طول- اریب، افراد زیادی تحقیق نموده اند که از آن جمله می‌توان به باتاچاریا و همکاران (۱۹۸۹)،

---

۱ Length-biased

جونز (۱۹۹۱)، گیامون و همکاران (۱۹۹۸)، افروموو یچ (۲۰۰۴) و همچنین چوبی و همکاران (۲۰۱۰) اشاره کرد. کریستوبال و آلکالا (۲۰۰۱) مرور خوبی بر تحقیقات انجام شده در زمینه داده‌های در طول-اریب انجام داده‌اند. فرض کنید $Y_1,\dots,Y_n$ نمونه‌ای تصادفی از توزیع G باشد. آنگاه با توجه به رابطه (۴)، یک برآوردگر طبیعی برای $f(x)$، به صورت

$$\widetilde{f}_n(x)=\frac{\widehat{\mu}_n\widehat{g}_n(x)}{x},\qquad x>0, \tag{۳}$$

است، که در آن

$$\widehat{g}_n(x)=\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{x-Y_j}{h}\right) \tag{۴}$$

برآوردگر هسته‌ای[۲] تابع چگالی g است که توسط روزن بلات (۱۹۵۶) معرفی شد. در (۴) تابع نامنفی $K(\cdot)$ را تابع هسته می‌نامند که در شرایط

$$\begin{aligned}
&i)\quad \int_{-\infty}^{\infty}K(x)dx=1,\\
&ii)\quad \int_{-\infty}^{\infty}xK(x)dx=0,\\
&iii)\quad \sigma^2=\int_{-\infty}^{\infty}x^2K(x)dx\neq 0
\end{aligned}$$

صدق می‌کند و h مقداری مثبت است که پهنای باند[۳] نامیده می‌شود. همچنین $\widehat{\mu}_n$ یک برآوردگر مناسب پارامتر μ است. با توجه به اینکه

$$E(\frac{1}{Y})=\int_0^{\infty}\frac{1}{y}g(y)dy=\frac{1}{\mu},$$

می‌توان μ را با روش گشتاوری به صورت

$$\widehat{\mu}_n=\frac{n}{\sum_{i=1}^{n}Y_i^{-1}}$$

برآورد کنیم. برآوردگر $\widetilde{f}_n(x)$ توسط باتاچاریا و همکاران (۱۹۸۸) معرفی و خواص حدی آن از قبیل سازگاری ضعیف و نرمال بودن مجانبی مورد بررسی قرار گرفت.

---
[۲] Kernel estimator

[۳] Bandwidth

مشکل اساسی این برآوردگر این است که وقتی $x \to 0$ آنگاه $\tilde{f}_n(x) \to \infty$. برای رفع این مشکل جونز (۱۹۹۱) برآوردگر

$$\widehat{f}_n(x) = (nh)^{-1}\widehat{\mu}_n \sum_{j=1}^{n} K\left(\frac{x-Y_j}{h}\right) Y_j^{-1}, \qquad (5)$$

را معرفی کرد، که در آن $K(\cdot)$ یک هسته متقارن است. جونز (۱۹۹۱)، ضمن بررسی خواص بهینه این برآوردگر نشان داد این برآوردگر یک تابع چگالی احتمال نیز است. بعلاوه در نزدیکی صفر مشکل برآوردگر باتاچاریا و همکاران (۱۹۸۸) را ندارد و رفتار بهتری از خود نشان می‌دهد. وی با استفاده از شاخص $MSE(\hat{f}_n(x)) = \int(\hat{f}_n(x) - f(x))^2dx$، h بهینه را به دست آورد. شکل کلی این برآوردگر برای توابع هسته دلخواه (متقارن یا نامتقارن)، به صورت زیر است

$$\hat{f}_h(x) = n^{-1}\widehat{\mu}_n \sum_{j=1}^{n} Y_j^{-1} K(x; Y_j, h). \qquad (6)$$

در برآورد تابع چگالی به روش هسته دو عامل مهم، تابع هسته و پارامتر پهنای باند نقش مهمی را ایفا می‌کنند. هر چند لازم به یادآوری است که انتخاب پارامتر پهنای باند بسیار مهمتر از انتخاب تابع هسته است. منابع متعددی در این زمینه وجود دارند که از آن جمله می‌توان به روزن بلات (۱۹۵۶)، پارزن (۱۹۶۲)، سیلورمن (۱۹۸۵) و واند و جونز (۱۹۹۵) اشاره نمود. بطور کلی توابع هسته را می‌توان به دو رده هسته‌های متقارن و نامتقارن تقسیم کرد، که از جمله هسته‌های متقارن می‌توان به هسته‌های باکس کار[۴]، گاوسی و اپانچنیکوف[۵] اشاره کرد. همچنین هسته‌های گاما، گاوسی وارون[۶] و عکس گاوسی وارون [۷] از جمله هسته‌های نامتقارن هستند.

معمولاً هسته‌های نامتقارن در برآورد تابع چگالی با تکیه‌گاه محدود مورد استفاده قرار می‌گیرند. زیرا هسته‌های متقارن در برآورد این نوع توابع چگالی اغلب خوب

[۴] Boxcar kernel

[۵] Epanechnikov kernel

[۶] Inverse Gussian kernel

[۷] Reciprocal inverse Gussian kernel

عمل نمی‌کنند و مشکل اریبی مرزی[۸] را بوجود می‌آورند (کولاسکرا و پاجت، ۲۰۰۶).

چون h میزان هموار‌ی برآوردگر را کنترل می‌کند به آن پارامتر هموار‌ی نیز گفته می‌شود. با انتخاب پهنای باند کوچک، برآورد کم هموار و با پهنای باند بزرگ‌تر، برآورد بیش هموار‌ی برای تابع چگالی به دست می‌آید. در بحث برآورد تابع چگالی، معمولاً پارامتر پهنای باند مجهول است و باید برآورد شود. چندین روش برای برآورد h وجود دارد، که از آن جمله می‌توان به روش‌های نزدیک‌ترین همسایگی[۹]، اعتبارسنجی متقابل کمترین توان‌های دوم[۱۰] و اعتبارسنجی متقابل اریب[۱۱] اشاره کرد.

در تمامی این روش‌های h به صورت یک پارامتر فراموضعی[۱۲] در نظر گرفته می‌شود. گن‌گوپادهایا و چه‌آنگ (۲۰۰۲) با در نظر گرفتن رهیافت بیزی به برآورد موضعی h پرداختند. برای داده‌های سانسور شده کولاسکرا و پاجت (۲۰۰۶) با در نظر گرفتن تابع زیان توان دوم خطا و انتخاب پیشین مناسب برای پهنای باند، برآوردگر $h(x)$ را به صورت میانگین توزیع پسین به دست آوردند.

در این مقاله، در بخش ۲ به بررسی انتخاب پهنای باند به روش بیزی در برآوردگر جونز پرداخته می‌شود. اثبات قضایای این بخش در پیوست ارائه می‌شوند . در بخش ۳ با مطالعه شبیه سازی دو برآورد بیزی و اعتبارسنجی متقابل کمترین توان‌های دوم برای پهنای باند در برآوردگر جونز مقایسه می‌شوند. بحث و نتیجه‌گیری در بخش ۴ ارائه می‌شود.

۲ برآورد بیزی پهنای باند

در این قسمت، با فرض اینکه پهنای نوار خود یک متغیر تصادفی است و داده‌ها در طول-اریب هستند، از برآوردگر جونز برای تخمین زدن تابع چگالی f استفاده

[۸] Boundary bias
[۹] Nearest neighbour
[۱۰] Least squares cross validation
[۱۱] Biased cross validation
[۱۲] Global

می‌شود. این مسأله را با انتخاب دو هسته گاوسی وارون و گاوسی که به ترتیب توابعی نامتقارن و متقارن هستند، بررسی نموده و h به روش بیزی برآورد می شود. فرض کنید $\pi(h)$ تابع چگالی پیشین برای h باشد. با در نظر گرفتن نمونه تصادفی Y_1، ...، Y_n از توزیع G، چگالی پسین h در نقطه x به صورت زیر است

$$\hat{\pi}(h|Y_1,\ldots,Y_n,x)=\frac{\hat{f}_h(x)\pi(h)}{\int \hat{f}_h(x)\pi(h)dh}. \qquad (7)$$

در قضیه زیر، پهنای باند برآوردگر جونز به صورت موضعی و بر اساس تابع هسته گاوسی وارون و چگالی پیشین گاما وارونه محاسبه می‌شود. شکل کلی تابع هسته گاوسی وارون به صورت زیر است

$$K(x;y,h)=\frac{1}{\sqrt{2\pi h x^3}}e^{-\frac{1}{2h}(x-y)^2/xy^2}, \qquad x,h>0,-\infty<y<\infty. \qquad (8)$$

قضیه ۱ : فرض کنید تابع هسته، گاوسی وارون و توزیع پیشین h، گامای وارون با پارامترهای α و β به صورت

$$\pi(h)=\frac{1}{\beta^\alpha\Gamma(\alpha)h^{\alpha+1}}e^{-(\frac{1}{\beta h})}, \qquad \alpha>0,\beta>0. \qquad (9)$$

باشد. با در نظر گرفتن تابع زیان توان دوم خطا، برآورد بیزی h که میانگین توزیع پسین است در نقطه x برابر است با

$$h_n^*=h_n(x)=E[(h|Y_1,\ldots,Y_n,x)]=\frac{\sum_{j=1}^n Y_j^{-1}(\beta_j^*)^{(\alpha^*-1)}}{(\alpha^*-1)\sum_{j=1}^n Y_j^{-1}(\beta_j^*)^{(\alpha^*)}}. \qquad (10)$$

برهان : با جایگذاری تابع هسته (۸) در رابطه (۶) و با توجه به (۷) توزیع پسین h به صورت

$$\hat{\pi}(h|Y_1,\ldots,Y_n,x) \quad = \quad \frac{\sum_{j=1}^n Y_j^{-1}.\frac{1}{\sqrt{2\pi h x^3}}e^{-\frac{(x-Y_j)^2}{2hxY_j^2}}.\frac{1}{\beta^\alpha\Gamma(\alpha)h^{\alpha+1}}e^{-\frac{1}{\beta h}}}{\int\sum_{j=1}^n Y_j^{-1}.\frac{1}{\sqrt{2\pi h x^3}}e^{-\frac{(x-Y_j)^2}{2hxY_j^2}}.\frac{1}{\beta^\alpha\Gamma(\alpha)h^{\alpha+1}}e^{-\frac{1}{\beta h}}dh}$$

$$= \quad \frac{\sum_{j=1}^{n}(Y_j^{-1}/h^{\alpha^*+1}e^{\frac{-1}{\beta_j^* h}})}{\sum_{j=1}^{n}Y_j^{-1}\int(e^{\frac{-1}{\beta_j^* h}}/h^{\alpha^*+1})dh}$$

$$= \quad \frac{\sum_{j=1}^{n}(Y_j^{-1}/h^{\alpha^*+1}e^{\frac{-1}{\beta_j^* h}})}{\Gamma(\alpha^*)\sum_{j=1}^{n}Y_j^{-1}(\beta_j^*)^{\alpha^*}},$$

به‌دست می‌آید، که در آن

$$\beta_j^* = [\frac{1}{\beta}+\frac{(x-Y_j)^2}{2xY_j^2}]^{-1}$$

$$\alpha^* = \alpha+\frac{1}{2}, \qquad \alpha > \frac{1}{2}.$$

برآورد بیزی h میانگین توزیع پسین به صورت زیر است.

$$\begin{aligned} h_n^* &= \int h.\frac{\sum_{j=1}^{n}(Y_j^{-1}/h^{\alpha^*+1})e^{-1/(h\beta_j^*)}}{\Gamma(\alpha^*)\sum_{j=1}^{n}Y_j^{-1}(\beta_j^*)^{\alpha^*}}dh \\ &= \frac{\sum_{j=1}^{n}Y_j^{-1}}{\Gamma(\alpha^*)\sum_{j=1}^{n}Y_j^{-1}(\beta_j^*)^{\alpha^*}}\int\frac{e^{-1/h\beta_j^*}}{h^{\alpha^*}}dh \\ &= \frac{\sum_{j=1}^{n}Y_j^{-1}(\beta_j^*)^{(\alpha^*-1)}}{(\alpha^*-1)\sum_{j=1}^{n}Y_j^{-1}(\beta_j^*)^{(\alpha^*)}}. \end{aligned}$$

قضیه ۲ : فرض کنید Y_1، ...، Y_n نمونه‌ای تصادفی از توزیع G، تابع چگالی f کراندار و $T = \inf\{t : 1 - F(t) > 0\}$ باشد. اگر به ازای هر $0 < x < T$، وقتی $\lim_{n\to\infty} h_n^* = 0$ آنگاه

$$|\hat{f}_{h_n^*}(x) - f(x)| \xrightarrow{a.s} 0.$$

برهان : با در نظر گرفتن G_n به عنوان تابع توزیع تجربی Y_i ها می‌توان تابع توزیع F را به صورت زیر برآورد کرد

$$F_n(t) = \widehat{\mu}_n\int_0^t y^{-1}dG_n(y).$$

بنابراین

$$\begin{aligned}\hat{f}_{h_n^*}(x) &= n^{-1}\widehat{\mu}_n \sum_{j=1}^{n} Y_j^{-1} K(x;Y_j,h_n^*) \\ &= \widehat{\mu}_n \int_0^\infty u^{-1} K(x;u;h_n^*)dG_n(u) \\ &= \int_0^\infty K(x;u;h_n^*)dF_n(u). \qquad (11)\end{aligned}$$

در نظر بگیرید

$$f_{h_n^*}(x) = \int_0^\infty K(x;u,h_n^*)dF(u).$$

بنابراین

$$\begin{aligned}|\hat{f}_{h_n^*}(x) - f(x)| &\leq |\hat{f}_{h_n^*}(x) - f_{h_n^*}(x)| + |f_{h_n^*}(x) - f(x)| \\ &=: J_1 + J_2. \qquad (12)\end{aligned}$$

حال داریم

$$\begin{aligned}J_1 &= \left|\int_0^\infty K(x;u,h_n^*)dF_n(u) - \int_0^\infty K(x;u,h_n^*)dF(u)\right| \\ &= \left|\int_0^\infty K(x;u,h_n^*)d[F_n(u) - F(u)]\right| \\ &= \left|\int_0^\infty [F_n(u) - F(u)]dK_u(x;u,h_n^*)\right| \\ &\leq \sup_{0<t\leq T}|F_n(t) - F(t)| \left|\int_0^\infty dK_u(x;u,h_n^*)\right|. \qquad (13)\end{aligned}$$

از طرفی بنا بر قضیه ۲ هوروات (۱۹۸۵)، F_n برآوردگر به‌طور یکنواخت سازگار قوی برای $F(t)$ است. بنابراین هرگاه $n \to \infty$ داریم

$$\sup_{0<t\leq T}|F_n(t) - F(t)| \xrightarrow{a.s} 0.$$

همچنین برای n های بزرگ

$$\left|\int_{\circ}^{\infty} dK_u(x;u,h_n^*)\right| = \frac{1}{\sqrt{2\pi h_n^* x^3}} e^{-\frac{1}{2h_n^* x}}.$$

بنابراین، وقتی $n \to \infty$، آنگاه $J_1 \to \circ$. از طرف دیگر، مشابه برهان قضیه ۲ کولاسکرا و پاجت (۲۰۰۶)، داریم $J_2 \to \circ$. پس اثبات کامل می‌شود. □

در ادامه پهنای باند برآوردگر جونز، بر اساس تابع هسته گاوسی و چگالی پیشین گاما وارونه با پارامترهای α و β به صورت

$$\tau(h) = \frac{2}{\Gamma(\alpha)\beta^\alpha}\frac{1}{h^{2\alpha+1}}\exp(-\frac{1}{\beta h^2}), \quad \alpha > \circ, \beta > \circ, h > \circ, \qquad (14)$$

محاسبه می‌شود.

قضیه ۳ : برای تابع هسته گاوسی، توزیع پیشین (۱۴) و تابع زیان توان دوم خطا، برآوردگر بیزی h در نقطه x عبارت است از

$$h_n^* = \frac{\Gamma(\alpha)\sum_{i=1}^n Y_i^{-1}\{1/(\beta(Y_i-x)^2+2)\}^\alpha}{\sqrt{2\beta}\Gamma(\alpha+\frac{1}{2})\sum_{i=1}^n Y_i^{-1}\{1/(\beta(Y_i-x)^2+2)\}^{\alpha+\frac{1}{2}}}. \qquad (15)$$

برهان : با در نظر گرفتن رابطه (۷) و جایگذاری تابع هسته گاوسی در رابطه (۶) داریم

$$\begin{aligned}\hat{\pi}(h|Y_1,\ldots,Y_n,x) &= \frac{\hat{f}_h(x)\pi(h)}{\int \hat{f}_h(x)\pi(h)dh} \\ &= \frac{\sum_{i=1}^n Y_i^{-1}\frac{1}{h^{2\alpha+2}}\exp\{-\frac{1}{h^2}\frac{(Y_i-x)^2}{2}+\frac{1}{\beta}\}}{\int_{\circ}^{\infty}\sum_{i=1}^n Y_i^{-1}\frac{1}{h^{2\alpha+2}}\exp\{-\frac{1}{h^2}\frac{(Y_i-x)^2}{2}+\frac{1}{\beta}\}dh}\end{aligned} \qquad (16)$$

فرض کنید $\mu_i = \frac{1}{2}(Y_i-x)^2+\frac{1}{\beta}$، طرف راست رابطه (۱۶) به صورت

$$= \frac{\sum_{i=1}^n Y_i^{-1}(\frac{1}{h^{2\alpha+2}})\exp\{-\frac{1}{h^2}\frac{(Y_i-x)^2}{2}+\frac{1}{\beta}\}}{\sum_{i=1}^n Y_i^{-1}\int_{\circ}^{\infty}\exp(-\frac{\mu_i}{h^2})(\frac{1}{h^{2\alpha+2}})dh} \qquad (17)$$

حاصل می‌شود. برای بررسی مخرج کسر (۱۷) فرض کنید

$$u = \frac{1}{h^2}$$

داریم

$$\begin{aligned}\sum_{i=1}^{n} Y_i^{-1} \int_0^{\infty} \exp(-\frac{\mu_i}{h^2}) \frac{1}{h^{2\alpha+2}} dh &= -\frac{1}{2} \sum_{i=1}^{n} Y_i^{-1} \int_0^{\infty} \exp(-\mu_i u) u^{\alpha-\frac{1}{2}} du \\ &= \frac{1}{2} \sum_{i=1}^{n} Y_i^{-1} \Gamma(\alpha + \frac{1}{2}) \mu_i^{-(\alpha+\frac{1}{2})} \\ &= \sum_{i=1}^{n} Y_i^{-1} \frac{\Gamma(\alpha+\frac{1}{2})}{2} \{\frac{(Y_i - x)^2}{2} + \frac{1}{\beta}\}^{-\alpha-\frac{1}{2}}\end{aligned}$$

بنابراین

$$\hat{\pi}(h|Y_1, \ldots, Y_1, x) = \frac{\sum_{i=1}^{n} Y_i^{-1} (\frac{1}{h^{2\alpha+2}}) \exp\{-\frac{1}{h^2} \frac{(Y_i-x)^2}{2} + \frac{1}{\beta}\}}{\sum_{i=1}^{n} Y_i^{-1} (\Gamma(\alpha+\frac{1}{2})/2) \{(Y_i - x)^2/2 + \frac{1}{\beta}\}^{-\alpha-\frac{1}{2}}}.$$

پهنای باند موضعی به‌صورت

$$h_n^*(x) = \frac{\int_0^{\infty} \sum_{i=1}^{n} Y_i^{-1} \exp(-\mu_i/h^2) (\frac{1}{h^{2\alpha+1}}) dh}{\sum_{i=1}^{n} Y_i^{-1} (\Gamma(\alpha+\frac{1}{2})/2) \{(Y_i - x)^2/2 + \frac{1}{\beta}\}^{-\alpha-\frac{1}{2}}}.$$

محاسبه می‌شود، که صورت کسر آن عبارت است از

$$\begin{aligned}\int_0^{\infty} \sum_{i=1}^{n} Y_i^{-1} \exp(\frac{-\mu_i}{h^2}) \frac{1}{h^{2\alpha+2}} h dh &= \frac{1}{2} \sum_{i=1}^{n} Y_i^{-1} \int_0^{\infty} \exp(-\mu_i u) u^{\alpha-1} du \\ &= \frac{1}{2} \sum_{i=1}^{n} Y_i^{-1} \Gamma(\alpha) \mu_i^{-\alpha} \\ &= \frac{\Gamma(\alpha)}{2} \sum_{i=1}^{n} Y_i^{-1} \{\frac{(Y_i - x)^2}{2} + \frac{1}{\beta}\}^{-\alpha}\end{aligned}$$

بنابراین

$$h_n^*(x) = \frac{\Gamma(\alpha) \sum_{i=1}^{n} Y_i^{-1} \{(Y_i - x)^2/2 + \frac{1}{\beta}\}^{-\alpha}}{\Gamma(\alpha+\frac{1}{2}) \sum_{i=1}^{n} Y_i^{-1} \{(Y_i - x)^2/2 + \frac{1}{\beta}\}^{-\alpha-\frac{1}{2}}}$$

$$= \quad \frac{\Gamma(\alpha)\sum_{i=1}^{n} Y_i^{-1}\{1/(\beta(Y_i-x)^2+2)\}^{\alpha}}{\sqrt{2\beta}\Gamma(\alpha+\frac{1}{2})\sum_{i=1}^{n} Y_i^{-1}\{1/\beta(Y_i-x)^2+2)\}^{\alpha+\frac{1}{2}}}.$$

تذکر ۱ : برای بررسی سازگاری قوی برآوردگر $\hat{f}_{h_n^*}$ با هسته گاوسی می‌توان مشابه قضیه ۲ عمل کرد.

۳ شبیه‌سازی

در این بخش، برای مقایسه برآورد پهنای باند به دو روش بیزی و اعتبارسنجی متقابل کمترین توان‌های دوم مطالعه‌ای شبیه سازی با دو هسته گاوسی و گاوسی وارون انجام می‌شود. در این شبیه سازی‌ها، توزیع f عضو خانواده‌های گاما به صورت

$$f(t) = \frac{t^{\gamma-1}e^{-\frac{t}{\theta}}}{\theta^{\gamma}\Gamma(\gamma)}, \gamma > 0, \theta > 0, t \geq 0,$$

و وایبول به صورت

$$f(t) = \frac{\gamma t^{\gamma-1}e^{-(\frac{t}{\theta})^{\gamma}}}{\theta^{\gamma}}, \gamma > 0, \theta > 0, t \geq 0,$$

در نظر گرفته می‌شود . طبق رابطه (۴)، اگر توزیع نااریب $Gamma(\gamma,\theta)$ باشد، آنگاه توزیع جامعه در طول-اریب $Gamma(\gamma+1,\theta)$ خواهد بود. به طریق مشابه اگر توزیع نااریب $Weibull(\gamma,\theta)$ باشد، آنگاه توزیع جامعه در طول-اریب گاما تعمیم یافته به صورت

$$f(t) = \frac{\gamma\theta^{-\gamma(1+\frac{1}{\gamma})}t^{\gamma(1+\frac{1}{\gamma})-1}e^{-(\frac{t}{\theta})^{\gamma}}}{\Gamma(1+\frac{1}{\gamma})}, \gamma > 0, \theta > 0, t \geq 0.$$

است. در محاسبه برآورد بیزی h، از توزیع‌های پیشین (۹) و (۱۴) به ترتیب برای هسته‌های گاوسی وارون و گاوسی استفاده می‌شود. در این شبیه سازی‌ها اثر پارامترهای توزیع پیشین روی برآوردگر $\hat{f}_{h^*}(t)$ بررسی شده است.

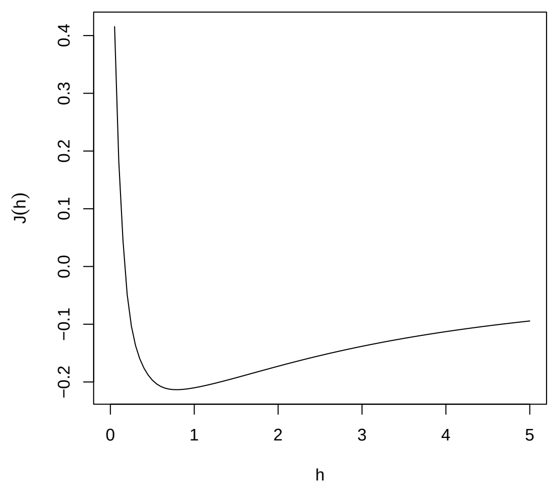
گیامون و همکاران (۱۹۹۸) نشان دادند در روش اعتبارسنجی متقابل کمترین توان‌های دوم، وقتی هسته متقارن است، h بهینه را می‌توان از کمینه کردن عبارت

$$\begin{aligned} J(h) &\cong ((\sum_{i=1}^{n} y_i^{-1})^{-2} h^{-1} \sum_{j=1}^{n}\sum_{i=1}^{n} (y_i y_j)^{-1} K*K(\frac{y_i-y_j}{h}) \\ &- 2\mu n^{-1} \sum_{i=1}^{n} (\sum_{j\neq i} y_j^{-1})^{-1} y_i^{-1} h^{-1} \sum_{j\neq i} y_j^{-1} K(\frac{y_j-y_i}{h})), \end{aligned}$$

به دست آورد، که در آن $K*K(\cdot)$ پیچش دو تابع K است و به صورت زیر تعریف می‌شود

$$K*K(u) = \int_{-\infty}^{\infty} K(u-v)K(v)dv.$$

نمودار $J(h)$ در شکل ۱، رسم شده است. به طریق مشابه، اگر هسته نامتقارن باشد،



شکل ۱: نمودار $J(h)$ در جامعه نااریب توزیع گاما با پارامترهای $\theta = 1, \gamma = 2$، $\alpha = \beta = 3$ و $n = 20$

می‌توان با جایگذاری رابطه (۶) در عبارت زیر و استفاده از روش اعتبارسنجی

متقابل کمترین توان‌های دوم، h بهینه را به دست آورد

$$\hat{J}(h) \cong \int \hat{f}_h^{۲}(t)dt - ۲\int \hat{f}_h(t)f(t)dt.$$

برای مقایسه عملکرد دو روش بیزی و اعتبارسنجی متقابل در برآورد h، نسبت MSE های برآورد شده برآوردگر جونز با دو پهنای باند به ازای مقادیر مختلف t به صورت

$$r(t) = \frac{EMSE(\hat{f}_{h_c}(t))}{EMSE(\hat{f}_{h_n^*}(t))},$$

محاسبه و نمودار آن در شکل‌های ۲ و ۴ رسم شده است، که در آن

$$EMSE(\phi_n(t)) = \frac{۱}{N}\sum_{i=۱}^{N}(\phi_n(t) - \phi(t))^{۲}.$$

$\phi(t)$ و $\phi_n(t)$ به ترتیب نشانگر تابع چگالی دلخواه و برآوردگر آن هستند و N تعداد دفعات شبیه سازی‌ها برابر ۱۰۰۰ در نظر گرفته شده است. h_c و h_n^* به ترتیب نشانگر مقدار h به دست آمده از روشهای اعتبارسنجی متقابل و بیزی هستند. بدیهی است که هر چقدر $EMSE(\hat{f}_{h_n^*}(t))$ کوچکتر از $EMSE(\hat{f}_{h_c}(t))$ باشد، روش بیزی از روش اعتبارسنجی متقابل بهتر عمل می‌کند. به عبارت دیگر برآوردگر به دست آمده با پهنای نوار بیزی بهتر از برآوردگر با پهنای نوار اعتبارسنجی متقابل است.

در شکل‌های ۲ و ۳ توزیع جامعه نااریب، گاما با پارامترهای $\theta = ۱$ و $\gamma = ۲$ بوده و هسته گاوسی است. همچنین پارامترهای توزیع پیشین h، $\alpha = \beta = ۳$ بوده و حجم نمونه $n = ۲۰$ است.

با توجه به شکل ۲، در نزدیکی مبدأ مختصات، ملاحظه می‌شود که روش بیزی در برآورد پهنای باند بهتر از روش اعتبارسنجی متقابل عمل می‌کند. اما به تدریج با زیاد شدن t، ابتدا روش اعتبارسنجی متقابل بهتر عمل کرده ولی با افزایش t هر دو روش تقریباً شبیه هم عمل می‌کنند.

شکل ۳، نمودار تابع چگالی جامعه نااریب گاما را به همراه برآوردگر جونز با پهنای باندهای برآورد شده به دو روش بیزی و اعتبارسنجی متقابل را نشان می‌دهد. همانگونه که در این شکل مشاهده می‌شود برآورد تابع چگالی به روش بیزی تا

نزدیکی نقطه ۲ نزدیک‌تر به تابع چگالی اصلی می باشد. بین نقاط ۲ تا ۵ برآورد تابع چگالی به روش اعتبارسنجی متقابل بهتر عمل می‌کند. از نقطه ۵ به بعد دو روش تقریباً شبیه هم عمل می‌کنند.

در شکل‌های ۴ و ۵ توزیع جامعه ناریب، وایبول با پارامترهای $\theta = ۱$ و $\gamma = ۲$ بوده و هسته گاوسی می‌باشد. همچنین پارامترهای توزیع پیشین h، $\alpha = \beta = ۳$ بوده و حجم نمونه $n = ۲۰$ است. با توجه به شکل ۴، در نزدیکی مبدأ مختصات، ملاحظه می‌شود که روش بیزی در برآورد پهنای باند بهتر از روش اعتبارسنجی متقابل عمل می‌کند. اما به تدریج با زیاد شدن t، روش اعتبارسنجی متقابل بهتر عمل می‌کند.

شکل ۵، نمودار تابع چگالی توزیع ناریب وایبول را به همراه نمودارهای برآوردگر جونز با پهنای باندهای برآورد شده به دو روش بیزی و اعتبارسنجی متقابل نشان می‌دهد. با توجه به این نمودار دو روش بیزی و اعتبارسنجی متقابل تقریباً شبیه هم عمل می‌کنند.

در ادامه ملاحظه می‌شود چگونه انتخاب پارامترهای توزیع جامعه ناریب (β و α) و توزیع پیشین h یعنی (θ و γ) و همچنین حجم نمونه‌گیری دارای نقش مهمی در ساختار برآوردگرهای جونز هستند.

در جدول ۱، مقدار جمع بسته میانگین توان دوم خطای برآورد شده برآوردگرها (EIMSE) با در نظر گرفتن هسته گاوسی و حجم‌های نمونه مختلف محاسبه شده است. در توزیع پیشین (۱۴)، برای مقادیر متفاوت β، α برابر ۳ در نظر گرفته شده است. این نسبت به صورت

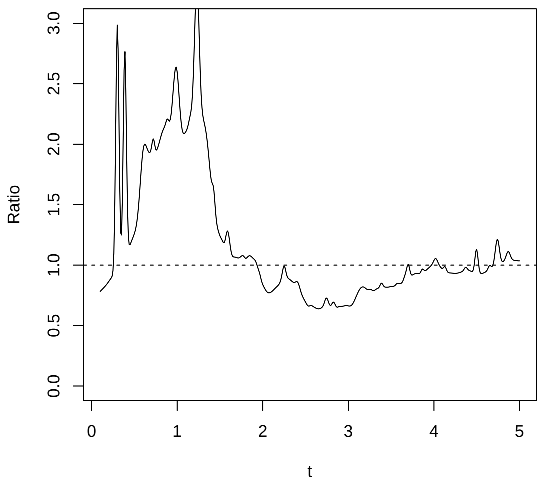
$$r^* = \frac{EIMSE(\hat{f}_{h_c})}{EIMSE(\hat{f}_{h^*})},$$

است، که در آن

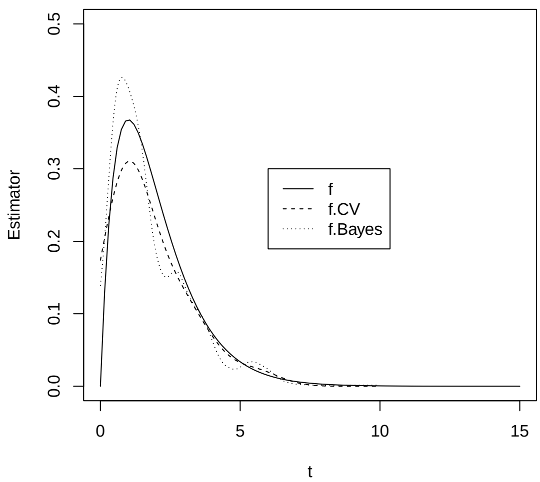
$$EIMSE(\phi_n) = \frac{۱}{mN}\sum_{j=۱}^{N}\sum_{i=۱}^{m}(\phi_n(z_i) - \phi(z_i))^۲.$$

برای محاسبه مقدار فوق، z_i ها به صورت

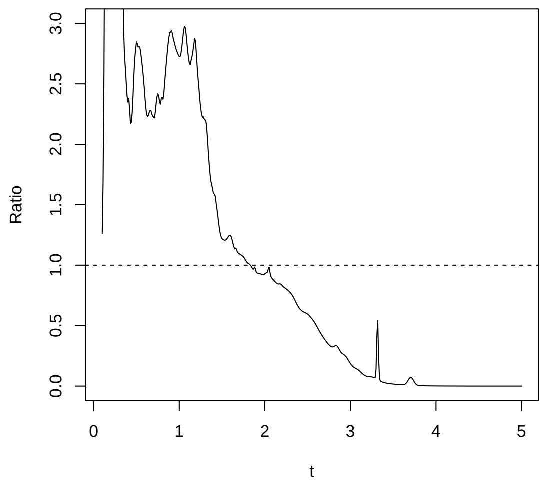
$$z_i = Y_{(۱)} + \frac{i}{m}(Y_{(n)} - Y_{(۱)}), \quad i = ۱, \ldots, m$$



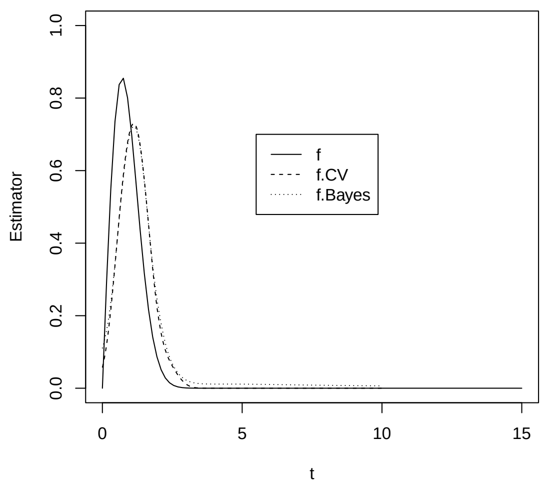
شکل ۲: نمودار $r(t)$ در خانواده گاما با پارامترهای $\gamma = ۲$، $\theta = ۱$، $\alpha = \beta = ۳$ و $n = ۲۰$



شکل ۳: نمودار تابع چگالی گاما و برآوردگر جونز با پهنای باندهای برآورد شده به دو روش بیزی و اعتبارسنجی متقابل



شکل ۴: نمودار $r(t)$ در خانواده ناریب وایبول با پارامترهای $\theta = ۱$، $\gamma = ۲$، $\alpha = \beta = ۳$ و $n = ۲۰$



شکل ۵: نمودار تابع چگالی وایبول و برآوردگر جونز با پهنای باندهای برآورد شده به دو روش بیزی و اعتبارسنجی متقابل

انتخاب می‌شوند، که در آن $Y_{(j)}$، j-امین آماره مرتب در نمونه تصادفی $Y_۱, \ldots, Y_n$ است.

جدول ۱: نسبت r^* برآوردگر جونز با پهنای باندهای اعتبارسنجی متقابل و بیزی در توزیع گاما و وایبول با $\theta = ۱$ و γ مختلف

$Weibull(۲, ۱)$	$Gamma(۲, ۱)$	$Exp(۱)$	$Weibull(۰/۵, ۱)$	$Gamma(۰/۵, ۱)$	β	n
۲/۹۴	۱/۹۹	۲/۷۵	۰/۸۰	۱/۹۸	۳	۵
۱/۵۹	۱/۳۲	۱/۹۰	۰/۵۵	۱/۵۲	۵	
۱/۰۵	۱/۲۱	۱/۷۵	۰/۲۹	۱/۰۷	۱۰	
۱/۶۶	۱/۴۳	۳/۳۵	۰/۹۸	۲/۱۱	۳	۲۰
۱/۷۴	۱/۰۱	۲/۶۷	۰/۸۳	۲/۱۶	۵	
۱/۲۶	۰/۶۶	۱/۷۳	۰/۶۸	۱/۹۱	۱۰	
۱/۸۵	۱/۵۶	۳/۲۱	۱/۰۵	۲/۱۶	۳	۴۰
۲/۳۸	۱/۱۵	۳/۰۹	۰/۹۱	۲/۴۹	۵	
۲/۰۹	۰/۷۷	۲/۸۵	۰/۷۶	۲/۶۱	۱۰	

همان‌طور که در جدول ۱ ملاحظه می شود که در تمامی حجم نمونه‌های انتخابی با افزایش پارامتر β مقدار نسبت r^* کاهش می‌یابد. لذا روش بیزی برای انتخاب پهنای باند در برآوردگر جونز با داده‌های در طول-اریب برای β های بزرگ بدتر از روش اعتبارسنجی متقابل عمل می‌کند. همچنین در خانواده وایبول در حجم نمونه پایین روش اعتبارسنجی متقابل بهتر از روش بیزی است.

۴ بحث و نتیجه‌گیری

در این مقاله روش بیزی برای انتخاب پهنای باند در برآوردگر جونز با داده‌های در طول-اریب استفاده شد و با روش اعتبارسنجی متقابل مقایسه گردید. قابل ذکر است که روش بیزی معایبی نیز دارد که از آن جمله می‌توان به مشکل انتخاب یک توزیع پیشین مناسب اشاره کرد. در اینجا انتخاب توزیع پیشین وابسته به نوع تابع هسته به کار رفته در محاسبه برآوردگر است.

مواردی وجود دارند که نیازمند تحقیق و بررسی بیشتر هستند و آن استفاده از توزیع‌های پیشین مناسب دیگر و هسته‌های متقارن و نامتقارن دیگر است. علاوه بر این همگرایی یکنواخت قوی برآوردگر تابع چگالی جونز و همگرایی میانگین مرتبه

دوم آن نیز برای خانواده هسته‌های متقارن و نامتقارن دیگر قابل بررسی است. علاقمندان برای دریافت برنامه‌های این مقاله که در محیط نرم افزار R تهیه شده‌اند می‌توانند با آدرس الکترونیک نویسنده مسئول مقاله تماس حاصل نمایند.

تقدیر و تشکر

نویسندگان از پیشنهادات داوران و هیئت تحریریه محترم مجله که باعث اصلاحات سازنده در این مقاله شده‌اند، کمال تشکر را دارند.

مراجع

Bhattacharyya, B. B., Franklin, L. A. and Richardson, G. D. (1988), A Comparison of Nonparametric Unweighted and Length-biased Density Estimation of Fibres, *Communications in Statistics-Theory and Methods*, **17**, 3629-3644.

Blumenthal, S. (1967), Proportional Sampling in Life Length Studies, *Technometrics*, **9**, 205-218.

Chaubey, Y. P., Sen, P. K. and Li, J. (2010), Smooth Density Estimation for Length-biased Data, *Journal of the Indian Society of Agricultural Statistics*, **64**, 145-155.

Cox, D. R. (1969), Some Sampling Problems in Technology, *New Developments in Survey Sampling*, Edited by Johnson and Smith, Wiley.

Cristobal, J. A. and Alcala, J. T. (2001), An Overview of Nonparametric Contributions to the Problem of Functional Estimation from Biased Data, *Sociedad de Estadistica e Investigacion Operativa*, **10**, 309-332.

Efromovich, S. (2004), Density Estimation for Biased Data, *The Annals of Statistics*, **32**, 1137-1161.

Gangopadhyay, A. K. and Cheung, K. N. (2002), Baysian Approach to the Choice of Smoothing Parameter in Kernel Density Estimation, *Nonparametric Statistics*, **14**, 655-664.

Guillamon, A., Navarro, j. and Ruiz, J. M. (1998), Kernel Density Estimation Using Weighted Data,*Communications in Statistics-Theory and Methods*, **27**, 2123-2135.

Horváth, L. (1985), Estimation From a Length-Biased Distribution, *Statistics and Decisions*, **3**, 91-113.

Jones, M. C. (1991), Kernel Density Estimation for Length Biased Data, *Biometrika* **78**, 511-519.

Kulasekera, K. B. and Padgett, W. J. (2006), Bayes Bandwidth Selection in Kernel Density Estimation with Censored Data,*Nonparametric Statistics*, **18**, 129-143.

Mcfadden, J. A. (1962), On the Lengths of Intervals in a Stationary Point Process, *Journal of the Royal Statistical Society*, **B 24**, 364-382.

Parzen, E. (1962), On Estimation of a Probability Density Function and Mode, *The Annals of Mathematical Statistics*, **33**, 1065-1076.

Rosenblatt, M. (1956), Remarks on Some Nonparametric Estimation of a Density Function, *The Annals of Mathematical Statistics*, **27**, 832-837.

Silverman, B. W. (1985), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Wand, M. P and Jones, M. C.(1985), *Kernel Smoothing*, Chapman and Hall, London.

Wicksell, S. D. (1925), The Corpuscle Problem. A Mathematical Study of a Biometrica Problem, *Biometrika*, **17**, 84-99.