پیش‌گفتار

برای مقایسه کردن توزیع ها آزمون مبتنی بر گشتاورها مانند آزمون میانگین ها یا واریانس ها و یا چندک ها مفید هستند و به خاطر مقایسه کردن یک تعداد از گشتاورها، اگرچه در وقت صرفه جویی می شود اما این نوع آزمون ناسازگار است. همچنین بکاربردن آزمون های پارامتری برای تساوی دو تابع چگالی این ضرورت را ایجاب می کند که توزیع صفر مشخص باشد و در غیر این صورت نتایج اشتباهی بدست خواهد آمد. بنابراین با وجود این شرایط، آزمون های ناپارامتری وجود دارند که می توانند کار مقایسه دو تابع توزیع بدون مشکلات ذکر شده در بالا انجام دهد.آزمون های مبتنی بر هسته برای آزمون تساوی توابع چگالی در حالت متغیرهای پیوسته بررسی شده که به احمد—ون بل (۱۹۷۴)، مامن (۱۹۹۲)، فن و جنسی (۱۹۹۳)، لی (۱۹۹۶) می توان مراجعه کرد. همچنین روش هسته مبتنی بر فراوانی یک برآورد سازگار برای تابع چگالی احتمال توام با داده های آمیخته پیوسته و گسسته است که با استفاده از این برآوردگر می توانیم به آسانی آزمون مبتنی بر هسته برای تساوی دو تابع چگالی مجهول بسازیم. در این روش هموارسازی روی متغیر گسسته انجام نمی شود در صورتی که ایتچیسون و ایتکن (۱۹۷۶)، هال (۱۹۸۱)، گراند و هال (۱۹۹۳)، اسکات (۱۹۹۲)، سمینف (۱۹۹۶)، لی و راسین (۲۰۰۳) و هال و همکاران (۲۰۰۴و۲۰۰۷) روی هموارسازی متغیر گسسته و مزیت های آن بحث کردند. بنابراین آزمون های ناپارامتری مورد بررسی قرار گرفتند که متغیر گسسته را نیز هموار کند که این کار توسط لی و همکاران در (۲۰۰۹) انجام شد و در این پایان نامه این روش را بررسی می کنیم. اساس این کار بر اساس برآورد هسته تابع چگالی احتمال می باشد که در آن هموارسازی هم روی متغیرهای پیوسته و هم روی متغیرهای گسسته انجام می شود. روش برآورد هسته ناپارامتری برای برآورد توزیع توام روی داده های چندگانه دودویی اولین بار توسط ایتچیسون و ایتکن در (۱۹۷۶) پیشنهاد شد و همچنین روش مبتنی بر درستنمایی برای انتخاب پارامترهای هموارسازی پیشنهاد کردند. امادر شبیه سازی نشان داده می شود که این برآوردگر برای داده های توام پیوسته و گسسته که متغیر پیوسته آن از توزیع دم پهن آمده باشد ناسازگار است که دلیل آن استفاده از روش درستنمایی برای انتخاب پارامتر هموارسازی است. لی و راسین در (۲۰۰۳) روش

هسته ناپارامتری با داده های آمیخته پیوسته و گسسته پیشنهاد کردند که پارامتر هموارسازی آن از طریق کمترین مربعات اعتبار مقطعی بدست می آید و مشکل برآوردگر ایتچیسون و ایتکن را نداشت، در این پایان نامه این برآوردگر را بررسی می کنیم. این پایان نامه شامل ۴ فصل می باشد که خلاصه مطالب هر فصل به شرح زیر است.

- در فصل ۱، روش های مختلف برآورد ناپارامتری تابع چگالی احتمال با داده های پیوسته و در حالت تک متغیره را بیان می کنیم.
- در فصل ۲، برآورد تابع چگالی احتمال با داده های توام پیوسته و گسسته را بیان می کنیم.
- در فصل ۳، آزمون ناپارامتری تساوی دو تابع چگالی احتمال با داده های توام پیوسته و گسسته را بیان می کنیم.
- در فصل ۴، برآورد ناپارامتری از تابع چگالی احتمال شرطی را بیان و آزمون ناپارامتری برای تساوی دو تابع چگالی احتمال شرطی را مطرح می کنیم.

در این پایان نامه مطالبی که با (*) نشان داده شده، مطالبی است که در کتاب نامه آمده است ولی به اثبات نپرداخته است یا برنامه شبیه سازی موجود نبوده است.

معصومه اکبری لاکه

شهریور ۱۳۸۹

فهرست مندرجات

فصل ۱

روش‌های مختلف برآورد ناپارامتری تابع چگالی احتمال برای متغیرهای پیوسته یک بعدی

۱–۱ برآوردگر بافت نگار

۱–۲ برآوردگر فراوانی چند ضلعی

۱–۳ تغییر دادن پارامترهموارسازی h

۱–۴ برآوردگر هسته چگالی

۱–۵ مشکلات برآوردگر هسته چگالی

۱–۶ تعدیل‌ها و بهبودها در برآوردگر هسته چگالی احتمال

۱–۷ هسته‌های مرتبه بالاتر

۱–۸ برآورد مبتنی بر تبدیل

۱—۱ برآوردگر بافت نگار

یک مفهوم اساسی در آنالیز داده تک متغیره بحث تابع چگالی احتمال است. اگر X متغیر تصادفی با تابع چگالی احتمال $f(x)$ باشد با استفاده از آن می‌توانیم توزیع X را توصیف کنیم و همچنین احتمال‌هایی را با استفاده از روابط زیربدست آوریم

$$p(a < X < b) = \int_a^b f(u)du$$

یک انگیزه برای ساختن برآورد ناپارامتری تابع چگالی احتمال با استفاده از تعریف تابع چگالی احتمال است

$$f(x) = \frac{d}{dx}F(x) = lim_{h \longrightarrow ۰}\frac{F(x+h) - F(x)}{h}$$

۱—۱—۱ برآوردگر بافت نگار

اگر $x_n, ..., x_۱$ یک نمونه تصادفی به حجم n از تابع چگالی احتمال f باشد و $F(x)$ را با تابع تو زیع تجربی، $\hat{F}(x) = \frac{\#\{x_i \leq x\}}{n}$ ، جایگذاری کنیم برآوردگر بافت نگار بصورت زیر خواهد بود.

$$\hat{f}(x) = \frac{(\#\{x_i \leq b_{j+۱}\} - \#\{x_i \leq b_j\})/n}{h} \qquad x\epsilon(b_j, b_{j+۱}]$$

یا

$$\hat{f}(x) = \frac{n_j}{nh} \qquad x\epsilon(b_j, b_{j+۱}] \qquad (۱—۱)$$

$b_n, ..., b_۱$، بیانگر باند است و n_j، تعداد مشاهداتی است که در باند jام قرار می‌گیرد و $h = b_{j+۱} - b_j$ است.

بدون شک رایج ترین برآوردگر چگالی تک متغیره، بافت نگارها هستند زیرا محاسبه آسان و ساختمان و تفسیری ساده دارند و نیاز به ابزار پیشرفته گرافیکی هم نیست.

۱—۱—۲ ویژگی‌های بافت نگار

از رابطه (۱—۱) واضح است که ویژگی‌های برآوردگر بافت نگار به h یا به عبارت دیگر به تعداد باندها وابسته است. در شکل ۱—۱ نمودارهای c, b, a بافت نگارهایی با ۲۰ باند هستند که هر سه بطور تصادفی از توزیع نرمال استاندارد به حجم ۱۰۰ تولید شده‌اند در حالی که نمودارهای d, e, f بافت نگارهایی با ۴ باند هستند و چگالی نرمال واقعی بر روی هر نمودار نیز رسم شده است.
نمودارهای c, b, a ناهمواری زیادی را نشان می‌دهد. در این نمودارها اگرچه ارتفاع باندها از چگالی واقعی پیروی می کند ولی آنها بطور قابل توجهی از یک نمودار به نمودار دیگر تغییر می‌کنند یعنی برآوردگر حاصل اریبی کم و تغییرپذیری زیاد دارد. در مقایسه، نمودارهای d, e, f هموار زیادی را نشان می‌دهند. در این نمودارها ارتفاع باندها بطور معمولی از یک نمودار به نمودار دیگر ثابت است اما آنها خیلی خوب از چگالی واقعی پیروی نمی‌کنند یعنی برآوردگر تغییرپذیری کم و اریبی بزرگ دارد. یک روش برای ارزیابی $\hat{f}(x)$ از طریق اندازه تفاوت آن با $f(x)$ است و ساده ترین اندازه هم مربعات خطاست که بصورت زیر می‌باشد.

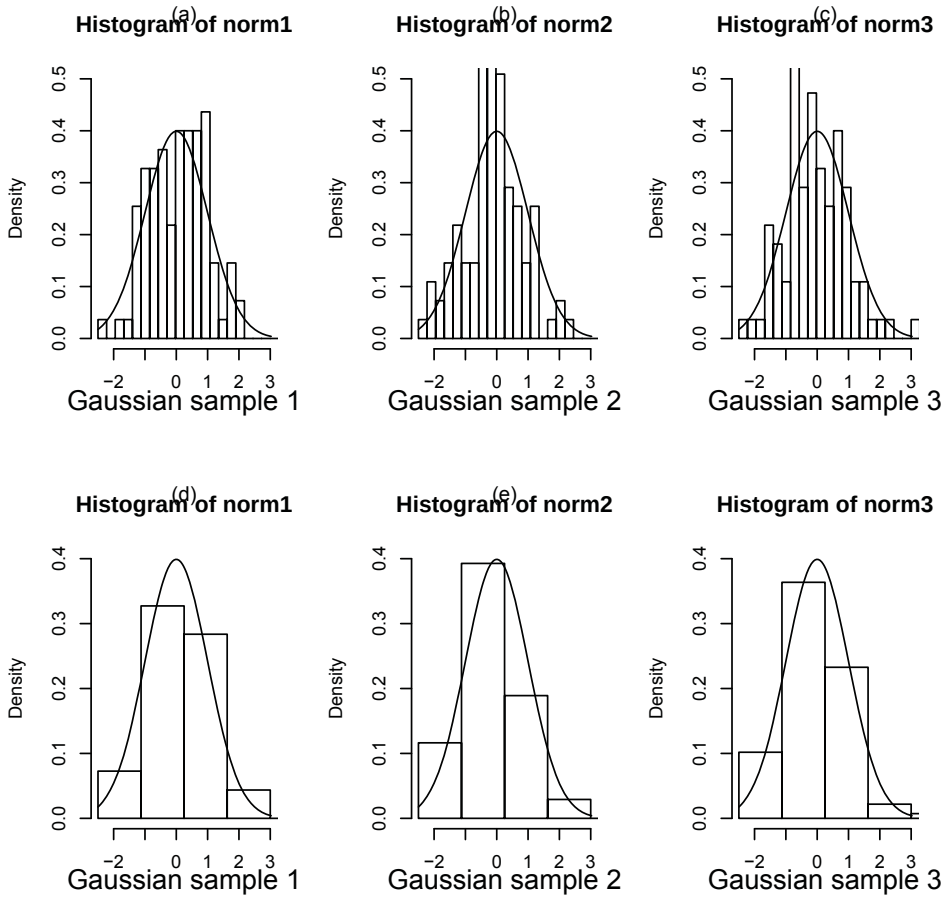
$$SE(x)^{۱} = [\hat{f}(x) - f(x)]^{۲}$$

$$MSE^{۲} = E_f[\hat{f(x)} - f(x)]^{۲}$$

$$ISE^{۳} = \int_{-\infty}^{\infty} [\hat{f}(u) - f(u)]^{۲} dx$$

---

۱ *square error*
۲ *mean square error*
۳ *integral square error*



شکل ۱ـ۱: برآوردگر بافت نگار با داده‌های تصادفی نرمال

چگالی $f(x)$، را هموار گوییم هرگاه $f^{'}(x)$ مطلقاً پیوسته و توان دوم آن انتگرال پذیر باشد در اینصورت می‌توان نشان داد که MSE مجانبی برابر است با

$$Bias[\hat{f}(x)] = E_f[\hat{f}(x)] - f(x) = ۱/۲f^{'}(x)[h - ۲(x - b_j)] + O(h^۲) \qquad x\epsilon(b_j, b_{j+۱}]$$

$$var(\hat{f}(x)) = \frac{f(x)}{nh} + O(n^{-۱})$$

$$\begin{aligned} MSE(\hat{f}(x)) &= var(\hat{f}(x)) + [Bias(\hat{f}(x))]^{۲} \\ &= \frac{f(x)}{nh} + \frac{[f^{'}(x)]^{۲}}{۴}[h - ۲(x - b_j)]^{۲} + O(n^{-۱}) + O(h^{۳}) \quad (۲-۱) \end{aligned}$$

سرانجام با انتگرال گرفتن بر روی همه باندها [۴] $MISE$ به این صورت است

$$MISE = \frac{۱}{nh} + \frac{h^{۲}R(f^{'}(x))}{۱۲} + O(n^{-۱}) + O(h^{۳}) \quad (۳-۱)$$

که $R(g)$ بیانگر $\int g^{۲}(u)\ du$ است.

مبادله اریبی و واریانس که قبلاً توجه کردیم در رابطه (۱—۳) نیز دیده می‌شود. h بطور مستقیم وابسته به انتگرال توان دوم اریبی ($\frac{h^{۲}R(f^{'}(x))}{۱۲}$) هست و رابطه معکوس با انتگرال واریانس ($\frac{۱}{nh}$) دارد. یعنی باندهای تنگ تر منجر به برآوردگری می‌شود که اریبی آن کم تر است اما تغییر پذیری زیاد دارد در نتیجه ساختن باندها با پهنای بیش تر اریبی را زیاد می‌کند.

به آسانی قابل محاسبه است h ی که$MISE$ را مینیمم می‌کند برابر است با

$$h_O = [\frac{۶}{R(f^{'}(x))}]^{\frac{۱}{۳}} n^{-۱/۳} \quad (۴-۱)$$

با جایگذاری این h بهینه در رابطه (۱—۳) داریم

$$AMISE_O = [\frac{۹R(f^{'}(x))}{۱۶}]^{\frac{۱}{۳}} n^{\frac{-۲}{۳}} \quad (۵-۱)$$

رابطه(۱—۵) نشان می‌دهد که سرعت همگرایی برآوردگر بافت نگار تابعی از $n^{\frac{-۲}{۳}}$ است. در ضمن $AMISE$ مجانبی بودن $MISE$ را نشان می‌دهد.

[۴] میانگین انتگرال مربع خطا

۱—۱—۳ انتخاب پارامتر هموارسازی، h

معادله (۱—۴) یک روش برای انتخاب کردن h در برآوردگر بافت نگار با پهنای ثابت است اما این روش به چگالی f بستگی دارد و ساده ترین روش برای انتخاب h،آن است که یک چگالی مرجع f انتخاب کنیم و سپس دررابطه (۱—۵) جایگذاری کنیم.
اگر چگالی مرجع را نرمال انتخاب کنیم h ی که $AMISE$ را مینیمم می‌کند به صورت زیر است

$$h_O = ۳.۴۹۱\sigma n^{\frac{-۱}{۳}} \qquad (۱—۶)$$

در این روش نیاز است که σ را برآورد کنیم.
مانع اصلی درانتخاب h بر اساس (۱—۶) این است که توجیه نظری برای این روش وقتی که چگالی اصلی نرمال نباشد، وجود ندارد.
اگر اندازه ISE را دوباره در نظر بگیریم

$$\begin{aligned} ISE &= \int [f(u) - \hat{f}(u)]^۲ du = \int f^۲(u)du + \int \hat{f}^۲(u)du - ۲\int \hat{f}(u)f(u)du \\ &= R(f) + R(\hat{f}(u)) - ۲\int \hat{f}(u)f(u)du \qquad (۱—۷) \end{aligned}$$

$R(f)$ به $\hat{f}$ بستگی ندارد پس درانتخاب h، تأثیری ندارد و جمله آخر (۱—۷) به ظاهر $-۲E(\hat{f}(u))$ هست که باید برآورد شود.
یک روش برای انجام کار قاعده اعتبار مقطعی [۵] است.
اگر $\hat{f}_{-i}(x_i)$ برآوردگر بافت نگار f با حذف مشاهده i ام باشد آنگاه

$$E(\hat{f}_{-i}(x_i)) = E(\hat{f}(x))$$

یا

$$E(\frac{۱}{n}\sum_{i=۱}^{n} \hat{f}_{-i}(x_i)) = E(\hat{f}(x)) = \int \hat{f}(u)f(u)du$$

[۵] $Cross\ validation$

بنابراین یک روش برای انتخاب h، قاعده اعتبار مقطعی است.

$$CV = R(\hat{f}) - \frac{۲}{n}\sum_{i=۱}^{n}\hat{f}_{-i}(x_i)$$

۱—۲ برآوردگر فراوانی چند ضلعی

۱—۲—۱ ویژگی‌هایی از برآوردگر فراوانی چند ضلعی

اگرچه بافت نگارها مفید هستند اما بدلیل ناهمواریشان بدنبال برآوردگرهای هموارتری هستیم. یک روش برای ساختن برآوردگر هموارتر این است که نقاط میانی بازه‌ها را با خط‌های مستقیم به هم وصل کنیم.

این برآوردگر را برآوردگر فراوانی چند ضلعی [۶] می‌نامیم. اگرچه پیوسته است اما مشتقات آن در نقاط میانی باندها تعریف نشده است.

اگر باندها با طول ثابت h را در نظر بگیریم و $\{b_۰, ..., b_{k+۱}\}$ ابتدای باندها را نشان دهد و C_j ها را بصورت زیر تعریف کنیم

$$\begin{aligned} C_j &= \frac{b_j + b_{j+۱}}{۲} \qquad j = ۱, ..., k \\ C_{k+۱} &= b_{k+۱} + \frac{h}{۲} \\ C_۰ &= b_۱ - \frac{h}{۲} \end{aligned}$$

آنگاه برآوردگر فراوانی چند ضلعی بصورت زیر تعریف می‌شود

$$\hat{f}(x) = \frac{۱}{nh^۲}[n_j c_{j+۱} - n_{j+۱}c_j + (n_{j+۱} - n_j)x], \qquad x\epsilon[c_j, c_{j+۱}] \qquad (۸—۱)$$

که در آن $n_۰ = n_{k+۱} \equiv ۰$ است.

آنالیز مجانبی نشان می‌دهد که این بهبود ظاهری می‌تواند بهبود دقت برآوردگر را در برداشته باشد.

[۶] $frequency\ \ polygon$

فرض کنید مشتق دوم f مطلقاً پیوسته و $R(f')$ و $R(f'')$ متناهی باشند در این صورت $MISE$ خواهد شد

$$MISE = \frac{۲}{۳nh} + \frac{۴۹h^۴R(f'')}{۲۸۸۰} + O(n^{-۱}) + O(h^۶) \qquad (۹-۱)$$

جمله اول (واریانس) به طور معکوس با h و جمله دوم (توان دوم اریبی) بطور مستقیم با h تغییر می کند. تفاوت قابل ملاحظه رابطه (۱−۹) نسبت به رابطه (۱−۳) این است که توان دوم اریبی از $O(h^۲)$ به $O(h^۴)$ تبدیل شده است. یعنی اگر h را کوچک کنیم اریبی با سرعت بیشتری کم می‌شود.
با مینیمم کردن (۱−۹) h بهینه به صورت زیر بدست می آید

$$h_O = ۲[\frac{۱۵}{۴۹}R(f'')]^{\frac{۱}{۵}} n^{\frac{-۱}{۵}} \qquad (۱۰-۱)$$

پس

$$AMISE_O = \frac{۵}{۱۲}[\frac{۴۹R(f'')}{۱۵}]^{\frac{۱}{۵}} n^{\frac{-۴}{۵}}$$

مشاهده می‌کنیم $AMISE$ با سرعت $O(n^{\frac{-۴}{۵}})$ به صفر میل می‌کند که در مقایسه با $AMISE$ در برآوردگر بافت نگار بیشتر شده است. مقایسه کردن (۱−۱۰) و (۱−۴) نشان می‌دهد که h بهینه برای برآوردگر فراوانی چند ضلعی فرم متفاوت از بافت نگار دارد و بطور مجانبی بزرگتر می‌شود.

۱−۲−۲ انتخاب پارامتر هموارسازی، h

ساده ترین روش برای انتخاب h برای برآوردگر فراوانی چند ضلعی جایگذاری یک فرم ویژه از f در (۱−۱۰) هست. اگر f رانرمال بگیریم، h بهینه به این صورت می شود

$$h_O = ۲.۱۵\sigma n^{\frac{-۱}{۵}}$$

۱—۳ تغییر دادن پارامتر هموارسازی، h

تا اینجا برآوردگرهای بافت نگار و فراوانی چند ضلعی بر اساس h ثابت بررسی شدند. می دانیم که h مبادله اریبی و واریانس در هر نقطه را کنترل می‌کند وهم‌چنین این مبادله به ویژگی‌های موضعی چگالی وابسته است.

معادله (۱—۲) نشان می‌دهد که ویژگی‌های موضعی چگالی در هر نقطه x، دقت بافت نگار در x را تعیین می‌کند. با توجه به آنچه گفته شد h در مناطقی با چگالی زیاد باید بزرگتر باشد تا جمله‌اول یعنی واریانس در MSE کاهش یابد وباید بطور عکس نسبت به $|f^{'}(x)|$ تغییر کند تا جمله دوم اریبی مینیمم شود.

تعریف بافت نگار که h آن به طور موضعی تغییر می‌کند به صورت زیر است

$$\hat{f}(x) = \frac{n_j}{n(b_{j+1} - b_j)} \qquad x\epsilon(b_j, b_{j+1}] \qquad (۱۱—۱)$$

حال مسئله، پیدا کردن باندهای مناسب $b_1, ..., b_{k+1}$ برای برآوردگراست.

$MISE$ مجانبی برآوردگر در باندی که شامل x هست فرم زیر را دارد

$$AMISE(x) = \frac{۲f(x)}{۳nh} + \frac{۴۹h^۴ f''(x)^۲}{۲۸۸۰}$$

در نتیجه h موضعی بهینه عبارت است از

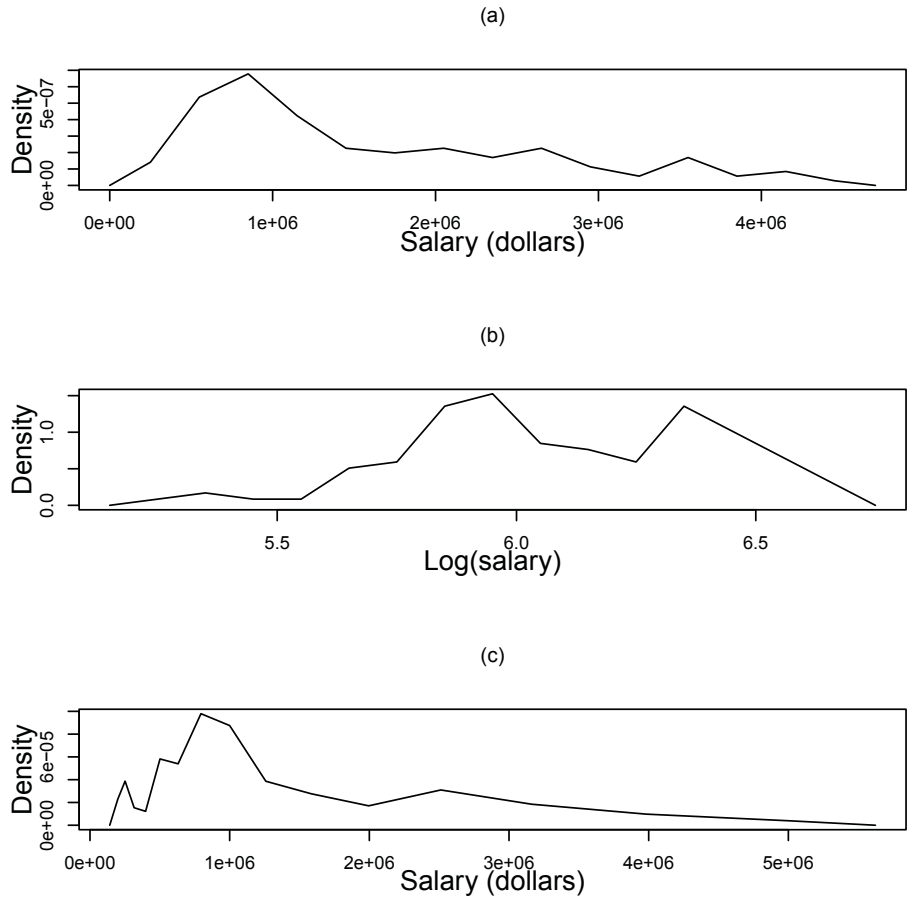
$$h_x = ۲[\frac{۱۵f(x)}{۴۹f''(x)^۲}]^{۱/۵} n^{-۱/۵}$$

برآوردگر فراوانی چند ضلعی با h موضعی به صورت زیر تعریف می‌شود

$$\hat{f}(x) = \frac{۱}{n(c_{j+1} - c_j)}[\frac{n_j c_{j+1}}{b_{j+1} - b_j} - \frac{n_{j+1} c_j}{b_{j+2} - b_{j+1}} + (\frac{n_{j+1}}{b_{j+2} - b_{j+1}} - \frac{n_j}{b_{j+1} - b_j})x] \quad x\epsilon[c_j, c_{j+1}]$$
$$(۱۲—۱)$$

یک روش ساده برای ساختن برآوردگرهای بافت نگار یا فراوانی چند ضلعی با h موضعی که در عمل هم خوب کار می کند، این است که ابتدا داده‌ها را به یک مقیاس دیگری تبدیل کنیم و سپس این

داده‌های تبدیل شده را هموار کنیم. سپس b_j های ساخته شده را به مقیاس اولیه برمی‌گردانیم و سپس از رابطه‌های (۱—۱۱) و (۱—۱۲) برای برآورد می‌توانیم استفاده کنیم. البته داده‌ها باید تحت تبدیل یکنوا تغییر مقیاس دهند.



شکل ۱—۲: برآوردگر فراوانی چندضلعی برای داده حقوق بازیکنان بیسبال

شکل ۱—۲ چگونگی روش انجام کار را شرح می‌دهد. داده‌ها مربوط به حقوق ۱۱۸ بازیکن بیسبال لیگ میجر در سال ۱۹۹۳ هستند.

در نمودار a، برآوردگر فراوانی چند ضلعی در مقیاس اصلی داده‌ها با h ثابت ۳۰۰۰۰۰ دلار محاسبه و رسم شده است. چنانکه شکل نشان می‌دهد دم سمت راست توزیع بلند است یعنی چوله به

راست است. بیش ترین مد حدود ۱ میلیون دلار هست. اما به دلیل ناهمواری در دم نمی‌توان ساختار دیگری در داده‌ها را تعیین کرد. بلند بودن دم، تبدیل لگاریتمی را پیشنهاد می‌کند.با این تبدیل ساختاری با عدم تاکید بلند بودن دم در شکل b قابل مشاهده است. در نمودار c، پس از تبدیل به مقیاس اصلی (با نمایی گرفتن) برآوردگر فراوانی چند ضلعی با h موضعی رسم شده است. اگرچه در این نمودار هنوز دم سمت راست بلند است اما علاوه بر مد ۱ میلیون دلاری، مدهای کوچکتری در ۲۵۰۰۰۰ دلار و ۵۰۰۰۰۰ دلار و ۲۵۰۰۰۰۰ دلار قابل مشاهده است.

۱—۴ برآوردگر هسته چگالی

۱—۴—۱ دلیل پیدایش برآوردگر هسته چگالی

اگرچه برآوردگرهای معرفی شده در بخش‌های قبلی مفید هستند اما دو اشکال عمده دارند، یکی اینکه هموار نیستند و دیگر آنکه به قدر کافی به ویژگی‌های محلی f حساس نیستند. برای این منظور تعریف تابع چگالی را به صورت زیر در نظر می‌گیریم

$$f(x) = \frac{dF(x)}{dx} = lim_{h \longrightarrow ۰} \frac{F(x+h) - F(x-h)}{۲h} \qquad (۱۳-۱)$$

بافت نگار (۱—۱۳) را با تقسیم کردن خط به باندها برآورد می‌کند اما یک روش منطقی این است که بطور جداگانه در هر نقطه x تابع چگالی را برآورد کنیم.
با جایگذاری تابع توزیع تجربی به جای $F(x)$ داریم

$$\hat{f}(x) = \frac{\#x_i \epsilon (x-h, x+h]}{۲nh}$$

که می‌تواند به این صورت هم نوشته شود

$$\hat{f}(x) = \frac{۱}{nh} \sum_{i=۱}^{n} k(\frac{x - x_i}{h}), \qquad (۱۴-۱)$$

که k بصورت زیر تعریف می‌شود

$$k(u) = \begin{cases} ۱/۲ & -۱ < u \leq ۱ \\ ۰ & otherwise. \end{cases}$$

فرم (۱—۱۴) برآوردگر هسته چگالی با تابع هسته k می‌باشد. توجه کنید که این تابع هسته، تابع چگالی یکنواخت روی [۱و۱—) هست.

یک تابع هسته هموارتر به برآورد هسته چگالی هموارتر منجر خواهد شد. شکل ۱—۳ برآوردگر هسته برای داده‌های نرخ سپرده سه ماهه ٦۹ بانک ایسلند با استفاده از چگالی نرمال برای k و h های ۰.۰۸ و ۰.۰۴ و ۰.۱٦ رسم شده است. فقط به‌ازای $h = ۰.۸$ منحنی بطور مطلوبی هموار شده است و یک فرم ۳ مدی با مدهای ۷/۵ و ۸ و ۸/۵ درصد نشان می‌دهد.

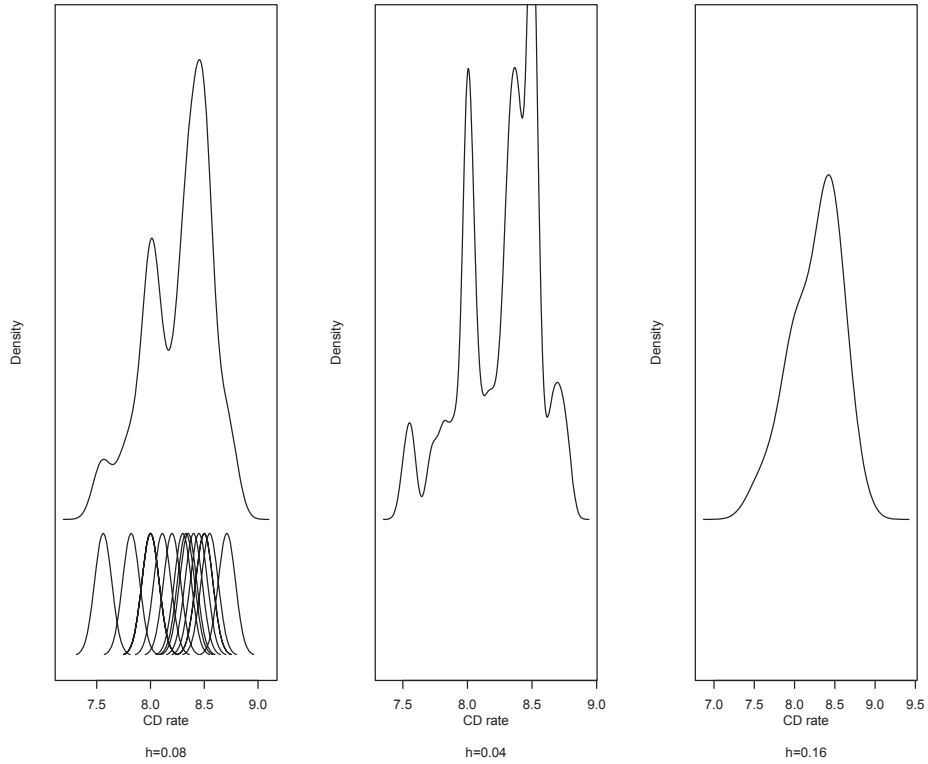
۱—۴—۲ چند ویژگی برآوردگر هسته

درجه هموارسازی داده ها تأثیر قوی روی ظاهر $\hat{f}(x)$ دارد و از طریق تنظیم پارامتر هموارسازی h قابل کنترل است.

انتخاب پارامتر هموارسازی می‌تواند از طریق معیاری مانند $MISE$ تعیین شود.

اگر $\hat{f}$ برآوردگر هسته چگالی باشد و تابع هسته k در شرایط زیر صدق کند آنگاه MSE، برآوردگر هسته بصورت زیر خواهد بود

$$\int k(u)du = ۱, \qquad \int uk(u)du = ۰, \qquad \int u^۲k(u)du = \sigma_k^۲ > ۰$$



شکل ۱ — ۳: برآوردگر هسته داده‌های نرخ سپرده سه ماهه ٦٩ بانک ایسلند

فرض کنید که چگالی اصلی به قدر کافی هموار باشد یعنی f'' بطور مطلق پیوسته و f'''^2 انتگرال پذیر باشد. اگر $h \longrightarrow 0$ و $nh \longrightarrow \infty$ وقتی که $n \longrightarrow \infty$ می‌رود، سپس با بسط سری تیلور داریم