دانشکده علوم ریاضی

گروه آمار

پایان‌نامه کارشناسی ارشد

رشته آمار ریاضی

عنوان

برآورد تابع چگالی به روش لگ-اسپلاین

استاد راهنما

دکتر حسینعلی نیرومند

استاد مشاور

دکتر حسن دوستی

تدوین

مهدیه عرفانیان

شهریور ۱۳۸۸

یا هو

چراغ، جز به نور چراغ، نتوان دید.

بسمه تعالی .

مشخصات رساله/پایان نامه تحصیلی دانشجویان .

دانشگاه فردوسی مشهد

عنوان رساله/پایان نامه: برآورد تابع چگالی به روش لگ-اسپلاین

نام نویسنده: مهدیه عرفانیان

نام استاد(ان) راهنما: دکتر حسینعلی نیرومند

نام استاد(ان) مشاور: دکتر حسن دوستی

دانشکده : علوم ریاضی	**گروه: آمار**	**رشته تحصیلی: آمار ریاضی**
تاریخ تصویب:		**تاریخ دفاع: ۱۳۸۸/۶/۲۸**
مقطع تحصیلی: کارشناسی ارشد ●	**دکتری ○**	**تعداد صفحات: ۱۴۸ صفحه**

چکیده رساله/پایان نامه :

در این پایان‌نامه پس از مرور تعدادی از روش‌های ناپارامتری برآورد چگالی که تاکنون ارائه شده‌اند، به معرفی اسپلاین‌ها، B-اسپلاین‌ها و روش بهینه‌سازی نیوتن-رافسون پرداخته‌ایم. مدل‌های خطی گسترش‌یافته را بیان و حالت خاص آن، یعنی مدل لگ-اسپلاین برای تابع چگالی را معرفی کرده‌ایم. در روش لگ-اسپلاین، ابتدا فرض می‌کنیم تابع چگالی جامعه به شکل تبدیل لجستیک است و سپس برآورد پارامترهای مجهول مدل را با استفاده از برآوردگر درست‌نمایی ماکزیمم به دست می‌آوریم که از روش نیوتن-رافسون محاسبه می‌شوند. به منظور به دست آوردن تعداد بهینه‌ی توابع پایه، از دو روش انتخاب گام به گام گره‌ها و ارائه‌ی برآوردگرهای سازگار ضعیف و قوی برای تعداد توابع پایه استفاده می‌کنیم. روش گام به گام از ترکیب دو روش اضافه کردن گره‌ها و حذف کردن گره‌ها تشکیل شده است که در آن‌ها به ترتیب آماره‌های رائو و والد به کار گرفته شده‌اند. در روش ارائه‌ی برآوردگر سازگار، تعداد توابع پایه را به صورت یک پارامتر مجهول در نظر می‌گیریم و سپس برآوردگرهایی را برای این پارامتر با استفاده از یک جانشین برای فاصله‌ی کولبک-لیبلر به دست می‌آوریم. اثبات می‌شود که بر حسب نوع محک اطلاعی که در به دست آوردن این برآوردگرها استفاده می‌کنیم، برآوردگرها سازگار ضعیف یا قوی خواهند بود. در انتها با استفاده از بسته‌ی لگ-اسپلاین موجود در نرم‌افزار آماری R به شبیه‌سازی پرداخته‌ایم و روش‌های ارائه‌شده برای یافتن تعداد بهینه‌ی توابع پایه را مقایسه کرده‌ایم.

کلید واژه:

B-اسپلاین، اسپلاین، انتخاب گام به گام گره‌ها، برآورد ناپارامتری تابع چگالی، برآورد درست‌نمایی ماکزیمم، برآوردگرهای سازگار ضعیف و قوی، روش بهینه‌سازی نیوتن-رافسون، محک اطلاع، مدل‌های خطی گسترش‌یافته.

امضای استاد راهنما:

تاریخ:

بسمه تعالی .

Graduate Studies Thesis\Dissertation Information
Ferdowsi University of Mashhad

Title of Thesis\Dissertation: Estimation of the density function by the Logspline method		
Author: Mahdiyeh Erfaniyan **Supervisor(s): Dr. Hosseinali Niroumand** **Advisor(s): Dr. Hassan Doosti**		
Faculty: Mathematical Sciences	**Department: Statistics**	**Specialization: Mathematical Statistics**
Approval Date:	**Defence Date: 22-August-2009**	
M.Sc. ● **Ph.D.** ○	**Number of Pages: 148**	

Abstract:

In this dissertation, after a review of some nonparametric density estimation methods which has been provided until now, we have introduced splines, B-splines and the Newton-Raphson optimization method. We have proposed the extended linear models and a special case of that, called the Logspline model of the density function. In the Logspline method, we first suppose that the density function of the population can be written in the form of the logistic transformation. We estimate the unknown parameters of model by the maximum likelihood estimators calculated by the Newton-Raphson method. In order to find the optimum number of basis functions, we use two methods which are the stepwise knots selection and providing the weakly and strongly consistent estimators for the number of basis functions. In the stepwise knots selection method, we use a combination of the knot addition and the knot deletion in which the Rao and Wald statistics are applied, respectively. For providing the weakly and strongly consistent estimators, we treat the number of basis functions as an unknown parameter and find some estimators for it using a proxy of the Kullback-Liebler distance. It is proved that these estimators are weakly and strongly consistent estimators in the order of the information criteria we have used for finding them. At the end, we have performed some simulation using the Logspline package written in R statistical language for comparing the two introduced method for finding the number of basis functions.

Signature of Supervisor: **Date:**	**Key Words:** **B-splines, Extended Linear Models, Maximum Likelihood Estimation, Nonparametric Density Estimation, Splines, Stepwise Knot Selection, the Newton-Raphson Optimization Method, Information Criteria, Weakly and Strongly Consistent Estimators.**

تقدیم به

او که نشانم داد فرشتگان خدا فقط در آسمان ها نیستند،

هستند فرشتگانی که در زمینند و مثل ما از خاکند،

او که خود فرشته‌ای زمینی است،

تقدیم به مادرم...

فهرست مطالب

پیش‌گفتار

در دهه‌های گذشته شاهد تغییر باورنکردنی در نظریه‌ی آمار بوده‌ایم. پیشرفت عظیم در توان کامپیوترها، باعث به وجود آمدن شیوه‌های تابعی سازگاری شده که اکنون ابزار ضروری برای تحلیل پیشرفته‌ی داده‌ها است. حال که آماردانان از فرض‌های سخت‌گیرانه‌ای که جزء بی چون و چرای مدل‌های پارامتری کلاسیک هستند، رهایی یافته‌اند، از آنان انتظار می‌رود که نه تنها قادر به برگزیدن متغیرهای مهم در یک مدل باشند، بلکه بتوانند شکل تابعی وابستگیِ این متغیرها را نیز انتخاب کنند. برای این که بتوان در کاربرد نیز توفیق یافت، به ناچار هر شیوه‌ی جدیدی باید بین انعطاف‌پذیری و مسئله‌ی خطیر تعداد ابعاد، تعادل برقرار کند. اسپلاین‌های چندجمله‌ای، انعطاف‌پذیری مورد نیاز برای تحلیل پیشرفته‌ی داده‌ها را به ارمغان می‌آورند. در این میان یافتن برآورد چگالی جامعه‌ی مورد بحث، یکی از پرتقاضاترین مباحث، هم در زمینه‌ی آمار نظری و هم در مباحث کاربردی است.

در این پایان‌نامه، به معرفی یک روش ناپارامتری برای برآورد چگالی، با عنوان لگ–اسپلاین می‌پردازیم. این روش از دو ابزار اساسی در برآوردیابی استفاده می‌کند که عبارتند از اسپلاین‌ها و برآورد درست‌نمایی ماکزیمم. اسپلاین‌ها که در اصل یکی از ابزارهای معرفی‌شده در ریاضیات کاربردی هستند، اینک کاربرد فراوانی در زمینه‌ی علوم دیگر، به خصوص مهندسی یافته‌اند. اسپلاین‌ها، یکی از مفاهیمی هستند که در کنار موجک‌ها، سری‌های فوریه و دیگر توابع پایه در

کار تقریب توابع مورد استفاده قرار می‌گیرند و خواص مطلوب آن‌ها در این امر، بر همگان واضح است. لگ–اسپلاین از مفهوم دیگری به نام برآورد درست‌نمایی ماکزیمم بهره می‌برد. شاید بتوان گفت این روش برآورد، ملموس‌ترین روش برآوردیابی است، زیرا در آمار جزو اوّلین روش‌هایی است که برای برآورد از آن یاد می‌شود و هم چنین بررسی‌های بسیاری برای یافتن خواص آن انجام شده است. روش درست‌نمایی ماکزیمم، ویژگی مهم دیگری دارد و آن هم این که در ارائه‌ی طیف گسترده‌ای از آزمون فرضیه‌های آماری کلاسیک به کار می‌رود. این روش، به خاطر مفهوم ملموس و قابل درک آن، حتی در بین دیگر رشته‌های علوم نیز روشی پرکاربرد و پرطرفدار است؛ چنان که تا کنون کارهای تحقیقی زیادی بر اساس روش درست‌نمایی ماکزیمم انجام یافته و حتی در انجام آزمون فرضیه‌ها نیز، بر حسب زمینه‌ی مورد لزوم، انواع سازگاریافته‌ای از این روش ارائه شده است، مانند انواع درست‌نمایی شرطی، حاشیه‌ای، برآوردشده، نیمرخ، جریمه‌شده، تجربی و .... بنابراین لگ–اسپلاین را به خاطر استفاده از این دو ابزار اساسی، می‌توان از یک سو روشی سازگار و کارآمد و از سویی دیگر، روشی ملموس و قابل فهم، حتی برای کاربران دیگر رشته‌ها، دانست.

در فصل اوّل این پایان‌نامه، مختصری از معروف‌ترین روش‌های ناپارامتری برآورد چگالی را بیان و سپس در فصل دوم، مفاهیم مورد نیاز از قبیل اسپلاین‌ها و روش‌های عددی به کار رفته در این شیوه را معرفی خواهیم کرد. سپس در فصل سوم مدل لگ–اسپلاین را تعریف و در فصل چهارم دو روش را برای یافتن تعداد توابع پایه ارائه می‌کنیم. در پایان و در فصل پنجم، بر اساس یافته‌های فصول پیشین، به شبیه‌سازی می‌پردازیم.

فرصت موجود، مغتنم است تا از استاد راهنمای محترم خود، جناب آقای دکتر حسینعلی نیرومند تشکر به عمل آورم، که در طول انجام این پایان‌نامه، از مراتب بالای علمی ایشان، و هم چنین از فضایل والای اخلاقیشان بهره‌های بسیار بردم. همین طور از استاد مشاور این کار، جناب آقای دکتر حسن دوستی سپاسگزارم، که با صبوری بسیار، مرا در مراحل مختلف تدوین این کار، راهنمایی کردند.

از جناب آقای دکتر مهدی عمادی و جناب آقای دکتر محمّد آرشی برای قبول داوری این کار، بسیار متشکرم.

از دوست و هم‌کلاسی عزیز خود، خانم حکیمه مربی هروی، به خاطر کمک‌های بی‌دریغشان در کار شبیه‌سازی این پایان‌نامه عرض تشکر فراوان دارم.

مهدیه عرفانیان
Email: mahdiyeh.erfaniyan@gmail.com

فصل اول

مروری بر معمول‌ترین روش‌های ناپارامتری برآورد چگالی

۱-۱ مقدمه

در این فصل مروری کوتاه بر روش‌های ناپارامتری معمول برآورد چگالی یک متغیر تصادفی خواهیم داشت. در بعضی موارد، لازم است بیان کنیم که اصولاً چرا به برآورد چگالی علاقه‌مند هستیم. یک دلیل مهم آن است که لزوم برآورد چگالی، در عمل آن قدر به حد کافی احساس می‌شود که طیف گسترده‌ای از تحقیقات آماری به آن اختصاص یافته است. بر این اساس، در ابتدا به بیان دلایل ارائه‌ی برآورد چگالی می‌پردازیم و در ادامه، روش‌های ناپارامتری مختلف برآورد چگالی را، به طور کوتاه، معرفی می‌کنیم.

۱-۲ چرا برآورد چگالی؟

به طور کلّی، می‌توان سه زمینه را بیان کرد که در آن‌ها به برآورد چگالی احتیاج است. اوّل آن که، برآورد چگالی می‌تواند برای شناسایی پدیده‌های الگودار مهم باشد. برای مثال، اگر بدانیم متغیری

که تحت مطالعه است دم‌های پهن یا قلّه‌های بلند دارد، هر مدل از داده‌ها که متناظر با چنین متغیری باشد، باید بتواند یک چگالی با چنین ویژگی‌هایی تولید کند. دوم آن که، برای برآوردگری که در یک مطالعه استفاده می‌شود، اغلب ممکن است یک تحلیل مونت کارلو[1] لازم باشد. به طور سنّتی، تنها تعداد کمی از گشتاورهای برآوردگر، ثبت یا آماره آزمون‌هایی مانند کولموگوروف–اسمیرنوف[2] برای ارزیابی انحراف از توزیع نرمال فراهم می‌شوند. اما برآوردهای ناپارامتری چگالی تصویر کاملی از توزیع برآوردگر را مهیا می‌کنند و بنابراین راه مطلوب‌تری برای خلاصه‌کردن خروجی آزمایش مونت کارلو به نظر می‌رسند. در نهایت آن که، گاهی اوقات، مسئله‌ی موجود آن است که برآوردگرهای پارامتری، توزیع مجانبی دارند که به یک چگالیِ محاسبه‌شده در یک نقطه‌ی مشخص بستگی دارد. برای مثال، میانه‌ی X دارای واریانس $0.25n^{-1}f^{-2}(0)$ است که در آن $f(0)$ چگالی X در نقطه‌ی $x=0$ است. بنابراین هر آماره آزمونی که شامل میانه باشد، به برآورد مقدار $f(0)$ احتیاج دارد.

حال فرض کنید $f=f(x)$ نشان‌دهنده‌ی تابع چگالی پیوسته‌ی متغیر تصادفی X در هر نقطه‌ی x باشد و $x_1,\ldots,x_n$ مشاهدات حاصل از f باشند. دو روش کلّی برای برآورد f پیشنهاد شده است:

(الف) برآوردهای پارامتری: روش‌های پارامتری شکل مخصوصی برای f در نظر می‌گیرند، مثلاً چگالی نرمال،

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

[1] *Monte Carlo analysis*
[2] *Kolmogorov-Smirnov*

که در آن میانگین μ و واریانس σ^2 پارامترهای f هستند. برآورد f می‌تواند به صورت زیر نوشته شود:

$$\hat{f}(x) = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \hat{\mu}}{\hat{\sigma}} \right)^2 \right],$$

که در آن μ و σ به راحتی از داده‌ها به صورت زیر برآورد می‌شوند:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{p}x)^2 \quad \text{و} \quad \hat{\mu} = \bar{p}x = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

(ب) برآوردهای ناپارامتری: یکی از نواقص روش پارامتری نیازمندبودن آن به دانستن قید چگالی واقعی f است. در روش ناپارامتری، $f(x)$ مستقیماً و بدون در نظر گرفتن شکل آن برآورد می‌شود. بافت‌نگار، یکی از روش‌های برآورد ناپارامتری است و یکی از قدیمی‌ترین روش‌های برآورد چگالی به شمار می‌رود (وان رایزین[1] (۱۹۷۳) و اسکات[2] (۱۹۷۹)). اما با وجودی که بافت‌نگار روش مفیدی برای برآورد چگالی است، مشکلاتی مانند ناپیوستگی و ناهموار[3]بودن را با خود دارد. علاوه بر این، برای استفاده در دو متغیر یا بیش‌تر بسیار پیچیده است. به خاطر این نواقص، در سه دهه‌ی گذشته، برآوردگرهای ناپارامتری مختلفی به کمک برآوردهای هموار[4] $f(x)$ ارائه شده‌اند.

[1] *Van Ryzin*
[2] *Scott*
[3] *rough*
[4] *smooth*

۱-۳ برآورد ناپارامتری چگالی

هیچ روش منحصربه‌فردی برای انجام برآورد ناپارامتری چگالی وجود ندارد. ما در این فصل، به معرفی هشت روش برآورد ناپارامتری تابع چگالی می‌پردازیم.

۱-۳-۱ رویکرد بافت‌نگار موضعی[1]

برای فهم برخی تکنیک‌های برآورد چگالی که بعداً مورد بحث قرار می‌گیرند، کار را با شرایطی آغاز می‌کنیم که در آن X یک متغیر تصادفی گسسته است. فرض کنید یکی از مقادیری که این متغیر می‌تواند بگیرد، x باشد و هدف ما برآورد $f(x)$ از داده‌های x_i که $i=1,\ldots,n$ است. برآورد $f(x)$ در حالت گسسته، لزوماً، برآورد نسبت مقادیر x در جامعه‌ی X است. با توجه به داده‌های $x_n,\ldots,x_1$ یک برآورد ساده و معروف، عبارت از نسبت نمونه یعنی $\hat{f_1}(x)=n^*/n$ است که در آن n^* تعداد $x_n,\ldots,x_1$های برابر با x است.

اکنون، حالتی را در نظر می‌گیریم که در آن X یک متغیر تصادفی پیوسته است، یعنی احتمال برابر بودن x با x_i مساوی صفر است و $f(x)$ لازم است با میانگین‌گیری از x_iهایی برآورد شود که در یک بازه در اطراف x ، مثلاً $x\pm h/2$ قرار دارد که در آن h عرض بازه است. بنابراین برآوردگر تجربی چگالی یعنی $\hat{f}(x)$ می‌تواند به صورت $\hat{f_1}(x)=(nh)^{-1}\sum_{i=1}^{n} I\left(x-\frac{h}{2}\le x_i\le x+\frac{h}{2}\right)$ بیان شود که در آن $I(\mathscr{A})=1$ اگر $\mathscr{A}$ درست باشد، و در غیر صورت برابر صفر است. به صورت دیگر، می‌توانیم بنویسیم

[1] *local histogram approach*

$$\hat{f_1}(x) = \frac{1}{nh}\sum_{i=1}^{n} I\left(-1/2 \le \frac{x_i - x}{h} \le 1/2\right)$$
$$= \frac{1}{nh}\sum_{i=1}^{n} I\left(-1/2 \le \psi_i \le 1/2\right), \qquad (۱-۳-۱)$$

که در آن $\psi_i = (x_i - x)/h$.

توجه کنید که $\hat{f}(x)$ در (۱-۳-۱) فراوانی نسبی به ازای هر واحد در بازه‌ی $(x - h/2, x + h/2)$ است که نقطه‌ی میانی آن x است. بنابراین به نظر می‌آید برآوردگر (۱-۳-۱) تلاش می‌کند بافت‌نگاری بسازد که بر پایه‌ی مشاهدات موضعی نسبت به x بوده، در آن هر نقطه‌ی x مرکز بازه‌ی نمونه‌گیری است. عرض بازه، یعنی h ، مقداری را کنترل می‌کند که به وسیله‌ی آن، داده‌ها برای تولید برآورد (۱-۳-۱) هموار (میانگین‌گیری) می‌شوند. هم چنین $\hat{f_1}$ ،مانند آن چه در فیکس و هاجز[1] (۱۹۵۱) آمده، به عنوان برآوردگر «اوّلیه»[2] شناخته می‌شود.

واضح است که تابع نشانگر یا تابع وزن $I\left(-1/2 < \psi_i < 1/2\right)$ در (۱-۳-۱) به فاصله‌ی x_i از x بستگی دارد. اگر قدر مطلق این فاصله، کم‌تر از یا مساوی با *1/2* باشد، وزن برابر ۱ و در غیر این صورت برابر صفر است. علاوه بر این موارد، تابع وزن $I(\psi) = I\left(-1/2 < \psi < 1/2\right)$ چنان است که

$$\int_{-\infty}^{\infty} I(\psi)d\psi = \int_{-\infty}^{-1/2} I(\psi)d\psi + \int_{-1/2}^{1/2} I(\psi)d\psi + \int_{1/2}^{\infty} I(\psi)d\psi$$
$$= \int_{-1/2}^{1/2} I(\psi)d\psi = \int_{-1/2}^{1/2} d\psi = 1. \qquad (۱-۳-۲)$$

بنابراین

---

[1] *Fix and Hodges*
[2] *naive*

$$\int_{-\infty}^{\infty} \hat{f}_1(x)dx = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} I\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right)dx$$
$$= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} I\left(-\frac{1}{2} < \psi_i < \frac{1}{2}\right)d\psi_i = 1, \qquad (۱-۳-۳)$$

و برآورد چگالی به این دلیل مناسب است که نامنفی و دارای انتگرال برابر با یک است. یک ویژگی (۱-۳-۳) آن است که انتگرال روی x گرفته شده، زیرا می‌تواند فرض کند که x روی تمام برد X مقدار می‌گیرد.

۱-۳-۲ اقتباس صوری[1] $\hat{f}_1(x)$

فرض کنید $F(x) = P(X \le x)$ نشان‌دهنده‌ی تابع توزیع احتمال تجمعی X باشد. آن گاه تابع چگالی $f(x)$ به صورت زیر تعریف می‌شود:

$$f(x) = \frac{d}{dx}F(x) = \lim_{h\to 0}\frac{F\left(x+\frac{h}{2}\right) - F\left(x-\frac{h}{2}\right)}{h}$$
$$= \lim_{h\to 0}\frac{P\left(x-\frac{h}{2} < X < x+\frac{h}{2}\right)}{h}. \qquad (۱-۳-۴)$$

مسئله‌ی ما برآورد $f(x)$ بر اساس $x_1,\ldots,x_n$ است. برای این منظور، فرض کنیم h تابع مثبتی از n باشد که با $n\to\infty$ به صفر میل می‌کند، و $P\left(x-\frac{h}{2} < X < x+\frac{h}{2}\right)$ را به وسیله‌ی نسبت مشاهدات $x_1,\ldots,x_n$ که در بازه‌ی $\left(x-\frac{h}{2}, x+\frac{h}{2}\right)$ قرار می‌گیرند، برآورد می‌کنیم. بنابراین یک برآورد بدیهی و ساده‌ی $f(x)$ در (۱-۳-۴) به صورت زیر است:

[1] *formal derivation*

$$\begin{aligned}\hat{f}_2(x) &= \frac{1}{nh}\left[\#(x_1 \ldots x_n) \in \left(x - \tfrac{h}{2}, x + \tfrac{h}{2}\right)\right] \\ &= \frac{1}{nh}\left[\#\left(\frac{x_1 - x}{h} \ldots \frac{x_n - x}{h}\right) \in (-1/2, 1/2)\right] \\ &= \hat{f}_1(x),\end{aligned} \qquad (۵-۳-۱)$$

که مانند (۱-۳-۱) است که در آن $\#(\mathscr{A})$ نشان‌دهنده‌ی عملگر «تعداد عناصر موجود در $\mathscr{A}$» است. برآوردگر داده‌شده در (۱-۳-۵) را اوّلین بار فیکس و هاجز (۱۹۵۱) معرفی کردند.

۱-۳-۳ برآوردگر هسته‌ی روزنبلات-پارزن[1]

برآوردگر چگالی که توسط تابع نشانگر معرفی‌شده در (۱-۳-۱) به دست می‌آید، این ویژگی را دارد که انتگرالش برابر یک است، اما نقصی هم دارد که ناهمواربودنش است. هم چنین $\hat{f}_1(x)$ یک تابع پیوسته نیست، و پرش‌هایی در نقاط $x_i \pm h/2$ با مشتقات صفر در سایر نقاط دارد. این امر به برآوردها، ماهیتی پله‌ای می‌دهد، و شاید ما مجموعه‌ای از وزن‌های هموارتر را ترجیح دهیم. روزنبلات (۱۹۵۶) این مسئله را با جایگذاری یک تابع هسته‌ی مثبت K که در شرط زیر صدق می‌کند، حل کرد:

$$\int_{-\infty}^{\infty} K(\psi)\, d\psi = 1. \qquad (۶-۳-۱)$$

برآورد کلّیِ وی به صورت زیر است:

[1] *Rosenblatt-Parzen kernel estimator*

$$\hat{f}(x)=\hat{f}_3(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)=\frac{1}{nh}\sum_{i=1}^{n}K(\psi_i), \qquad (۷-۳-۱)$$

که در آن $\psi_i=h^{-1}(x_i-x)$ مانند (۱-۳-۱) است و h ، عرض پنجره[1] (هم چنین پارامتر هموارسازی یا پهنای نوار نیز نامیده می‌شود)، تابعی به اندازه‌ی n است و با $n\to\infty$ به صفر میل می‌کند.

تعدادی از ویژگی‌هایی که یک هسته باید دارا باشد می‌تواند از ماهیت تابع نشانگر استنباط شود. اولاً، برای مقادیر بزرگ $|\psi_i|$ (یعنی این که x_i به فاصله‌ی زیادی از x قرار می‌گیرد) باید $K(\psi_i)$ کوچک باشد، همین طور وزن‌های بسیار کوچک باید به چنین داده‌هایی در ساخت برآورد چگالی اختصاص یابند. به خصوص، چون وقتی $n\to\infty$ آن گاه $h\to 0$ ، نتیجه می‌شود که برای هر $x_i\neq x$ داریم $|\psi_i|\to\infty$ ، و بنابراین $K(-\infty)=K(\infty)=0$ که از رابطه‌ی (۱-۳-۶) حاصل می‌شود. این ویژگی، خاصیت صفر تابع نشانگر را دوباره به دستم می‌دهد، که در آن قسمت یک با داشتن $\int_{-\infty}^{\infty}K(\psi)d\psi=1$ نشان داده می‌شود. این امر منجر به جایگذاری یک مربع به مرکزیت x و به طول واحد با یک منحنی هموار می‌شود، هم چنین می‌توان آن را با مربعی به مرکزیت x به همان مساحت اما نه لزوماً با محمل کراندار جایگذاری کرد. علاوه بر این، چون این ویژگی‌ها همان ویژگی‌های تابع چگالی هستند، هسته‌ها اغلب همان توابع چگالی معروف در نظر گرفته می‌شوند، برای مثال، هسته‌ی نرمال استاندارد به صورت $K(\psi)=(2\pi)^{-1/2}\exp(-.5\psi^2)$ است. در این راستا، تابع نشانگر را می‌توان به صورت یک برآوردگر هسته با $K(\cdot)$ در نظر گرفت که روی $[-1/2,1/2]$ دارای چگالی یکنواخت است.

[1] *band width*