مجله علوم آماری، پاییز و زمستان ۱۳۸۸
جلد ۳، شماره ۲، ص ۱۹۹-۲۰۹

برآورد تابع چگالی در حضور داده‌های پرت

عباس مهدوی، مینا توحیدی
گروه آمار، دانشگاه شیراز

تاریخ دریافت: ۱۳۸۸/۱۰/۱۰ تاریخ آخرین بازنگری: ۱۳۸۹/۳/۱۶

چکیده : وجود مشاهدات پرت یکی از مهمترین موضوعات در استنباط آماری است. با توجه به این که این مشاهدات تاثیر زیادی بر روی مدل برازش شده و استنباط‌های مربوط به آن دارند، پیدا کردن روشی برای مشخص کردن اثر مشاهدات پرت ضروری است. هدف این مقاله بررسی تاثیر مشاهدات پرت بر روی برآورد تابع چگالی به روش هسته‌ای است. در این مقاله با استفاده از روش جستجوی پیشرو، به شناسایی مشاهدات پرت و تاثیر آن‌ها بر برآورد تابع چگالی به روش هسته‌ای پرداخته می‌شود.

واژه‌های کلیدی : برآورد تابع چگالی هسته‌ای، روش جستجوی پیشرو، پارامتر هموارساز.

۱ مقدمه

جستجوی پیشرو[۱] یک روش توانا برای پیدا کردن مشاهدات پرت و مشخص کردن اثر آن‌ها بر روی مدل برازش شده است. این روش ابتدا با تعیین یک زیرمجموعه کوچک از مشاهدات آغاز می‌شود، اندازه این زیرمجموعه به طور مکرر افزایش پیدا

° آدرس الکترونیک مسئول مقاله: مینا توحیدی، Mtowhidi@susc.ac.ir
° کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲G۰۷ و ۶۲G۰۵

۱ Forward Search

می‌کند و در هر مرحله تعدادی از پارامترهای مدل برآورد می‌شوند. مشاهدات به طریقی انتخاب می‌شوند که مشاهدات پرت در انتهای جستجو وارد مدل شوند. در پایان با بررسی نمودار توابعی از برآورد پارامترها اثر مشاهدات پرت آشکار می‌شود. این روش در سه مرحله انجام می‌شود، مرحله اول انتخاب زیر مجموعه اولیه، مرحله دوم افزودن مشاهدات در طول جستجو و مرحله سوم نمایش آماره‌های مورد نیاز در طول جستجو است. این روش دارای کارایی بالایی است زیرا که قابلیت شناسایی گروهی مشاهدات پرت را دارد و تحت تاثیر سرپوش گذاشتن[۲] قرار نمی‌گیرد به این دلیل که مشاهدات مرحله به مرحله مورد بررسی قرار می‌گیرند.

این ایده که مدل به روی زیرمجموعه‌هایی با اندازه‌های افزایشی برازش داده شود نخستین بار توسط هادی (۱۹۹۲)، اتکینسون (۱۹۹۴) در داده‌های چند متغیره و هادی و سیمونوف (۱۹۹۳) در رگرسیون مطرح شد. بعد از آن اتکینسون و ریانی (۲۰۰۰) کاربرد جستجوی پیشرو را در رگرسیون خطی و غیرخطی، تبدیل متغییر پاسخ و مدل‌های خطی تعمیم یافته بیان کردند. برتاسینی و واریبل (۲۰۰۷) در مدل آنالیز واریانس و کوین (۲۰۰۸) در آزمون نرمال بودن از این روش استفاده کردند (برای اطلاع بیشتر در مورد نحوه پیشرفت استفاده از روش جستجوی پیشرو در مطالعات آماری، به مقالهٔ اتکینسون و همکاران (۲۰۱۰) مراجعه شود).

در این مقاله تاثیر مشاهدات پرت بر برآورد تابع چگالی هسته‌ای[۳] بررسی می‌شود و با استفاده از روش جستجوی پیشرو، روشی برای شناسایی مشاهدات پرت ارائه خواهد شد. پس از شناسایی مشاهدات پرت، در صورتی که نسبت این مشاهدات به کل مشاهدات کوچک باشد، نشان داده می‌شود که حذف داده‌های پرت منجر به برآورد مناسب‌تری برای تابع چگالی خواهد شد.

۲ برآورد تابع چگالی به روش هسته‌ای

نمونه تصادفی $X_1, \cdots, X_n$ را از یک توزیع پیوسته با چگالی احتمال $f(x)$ در نظر بگیرید. برآورد تابع چگالی هسته‌ای یک متغیره به صورت

---

۲ Masking

۳ Kernel Density Estimator

$$\hat{f}(x,h) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x-X_i}{h})$$

است، که در آن $k(x)$ تابع هسته نامیده شده و به گونه‌ای انتخاب می‌شود که $\int_{-\infty}^{+\infty} k(x)dx = 1$ باشد. h یک عدد مثبت است که پارامتر هموارساز[۴] نامیده می‌شود و میزان همواری نمودار تابع چگالی را کنترل می‌کند. معمولاً $k(x)$ به صورت یک تابع چگالی متقارن انتخاب می‌شود، در این صورت $\hat{f}(x,h)$ هم یک تابع چگالی احتمال است که تمام ویژگی‌های پیوستگی و مشتق‌پذیری $k(x)$ را به ارث می‌برد. در این مقاله از تابع هسته‌ای نرمال برای برآورد تابع چگالی استفاده می‌کنیم که به صورت

$$k(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

تعریف می‌شود. روش‌های مختلفی برای برآورد پارامتر هموارساز ارائه شده است. در این مقاله از برآورد پارامتر هموارساز $h = 0/9An^{-\frac{1}{5}}$ که برای تابع هسته‌ای نرمال توسط سیلورمن (۱۹۸۶) ارائه شده است، استفاده می‌شود، که در آن $A = \min\{S, \frac{Q}{1/34}\}$، n حجم نمونه، S انحراف معیار و Q دامنه میان چارکی نمونه می‌باشند. برای اطلاعات بیشتر می‌توان به روزنبلات (۱۹۵۶)، پارزن (۱۹۶۲) و سیلورمن (۱۹۸۶) مراجعه نمود.

۳ برآورد تابع چگالی هسته‌ای و جستجوی پیشرو

در این قسمت با استفاده از جستجوی پیشرو روشی برای شناسایی مشاهدات پرت در مسئله برآورد تابع چگالی ارائه می‌شود. این روش در سه مرحله انجام می‌شود مرحله اول انتخاب زیر مجموعه اولیه، مرحله دوم افزودن مشاهدات در طول جستجو و مرحله سوم نمایش آماره‌های مورد نیاز در طول جستجو است.

مرحله ۱. انتخاب زیر مجموعه اولیه: جستجوی پیشرو نسبت به روشی که زیرمجموعه اولیه را انتخاب می‌کند حساس نیست، به شرطی که این زیرمجموعه

[۴] Smoothing Parameter

شامل مشاهدات پرت نباشد. این زیرمجموعه با انتخاب $m_0 = [\frac{n}{4}]$ از مشاهدات X_i که کمترین مقدار $|X_i - med(X_i)|$ را داشته باشند ساخته می‌شود، که در آن $med(x_i)$ نشان دهنده میانه مشاهدات است. این زیرمجموعه با $S^{(*)}$ نشان داده می‌شود که اندازه آن برابر جزء صحیح نصف اندازه مشاهدات است.

مرحله ۲. اضافه کردن مشاهدات در طول جستجو: در این مرحله مشاهده‌ای که به زیر مجموعه قبلی نزدیک‌تر است اضافه می‌شود. برای این کار فرض کنید زیرمجموعه $S^{(m)}$ با اندازهٔ m داده شده باشد. برای رفتن به مرحله بعد با استفاده از زیرمجموعه $S^{(m)}$ تابع چگالی را برآورد کرده و بر اساس آن، برای تمام مشاهدات مقدار $\hat{f}(x,h)$ محاسبه می‌شوند. سپس $m+1$ مشاهده‌ای که بیشترین مقدار $\hat{f}(x,h)$ را دارند انتخاب می‌گردد. این کار ادامه می‌یابد تا تمام n مشاهده انتخاب شوند. به این نکته توجه داشته باشید که در اکثر مواقع وقتی که از زیرمجموعه m عضوی به زیرمجموعه $m+1$ عضوی می‌رسیم، فقط یک عضو اضافه می‌شود گرچه می‌تواند دو عضو یا بیشتر اضافه شده و یک عضو یا بیشتر حذف شود. با توجه به روش استفاده شده برای ورود داده‌های جدید، مشاهدات پرت در انتهای جستجو وارد می‌شوند و مسئله سرپوش گذاشتن بر داده‌های پرت ایجاد نمی‌گردد.

مرحله ۳. نمایش نتایج جستجوی پیشرو: در هر مرحله از جستجوی پیشرو آماره‌های مورد نیاز برای مشخص کردن مشاهدات پرت و آنالیز اثر آن‌ها روی برآورد تابع چگالی نمایش داده می‌شوند. در این‌جا در هر مرحله از جستجو، واریانس زیرمجموعه‌ای که با استفاده از آن تابع چگالی برآورد شده است، به دست آورده می‌شود. از آن‌جایی که در هر مرحله مشاهداتی انتخاب می‌شوند که بیشترین مقدار جرم احتمال را داشته باشند، واریانس مشاهدات در هر مرحله روند صعودی دارد. ولی با این وجود وقتی که مشاهدات پرت وارد شوند، در نمودار واریانس به طور ناگهانی سرعت صعود افزایش می‌یابد.

۴ شبیه‌سازی

به منظور ارزیابی این روش نمونه‌هایی از توزیع‌های مختلف مورد بررسی قرار می‌گیرد. با استفاده از جستجوی پیشرو رفتار مشاهداتی که عاری از مشاهدات پرت هستند با مشاهداتی که دارای آلودگی هستند مقایسه می‌شوند. در ابتدا شش نمونه به حجم ۱۰۰ که به طریق زیر ساخته می‌شوند را در نظر بگیرید:

نمونه A : ۱۰۰ مشاهده از توزیع نرمال استاندارد.

نمونه B : ۱۰۰ مشاهده از توزیع $t_{(df=۲)}$.

نمونه C : ۱۰۰ مشاهده از توزیع $\chi^{۲}_{(df=۲)}$.

نمونه D : ۱۰۰ مشاهده از توزیع $LogNorm(\mu = ۰, \sigma = ۱)$.

نمونه E : ۱۰۰ مشاهده از توزیع $Weibull(sh = ۲, sc = ۱)$.

نمونه F : ۱۰۰ مشاهده از توزیع $Beta(\alpha = ۲, \beta = ۲)$.

شکل ۱ نتایج جستجوی پیشرو را برای شش نمونه A تا F نشان می‌دهد. محور افقی در این نمودارها، مراحل باقیمانده تا وارد شدن تمام مشاهدات و محور عمودی واریانس مشاهدات تحت بررسی در هر مرحله را نشان می‌دهند. برآورد واریانس در این نمودارها روند صعودی دارد، اما هیچ تغییر ناگهانی در مراحل پایانی این نمودارها دیده نمی‌شود. پس با توجه به این نمودارها در این نمونه‌ها مشاهدهٔ پرت مهمی وجود ندارد.

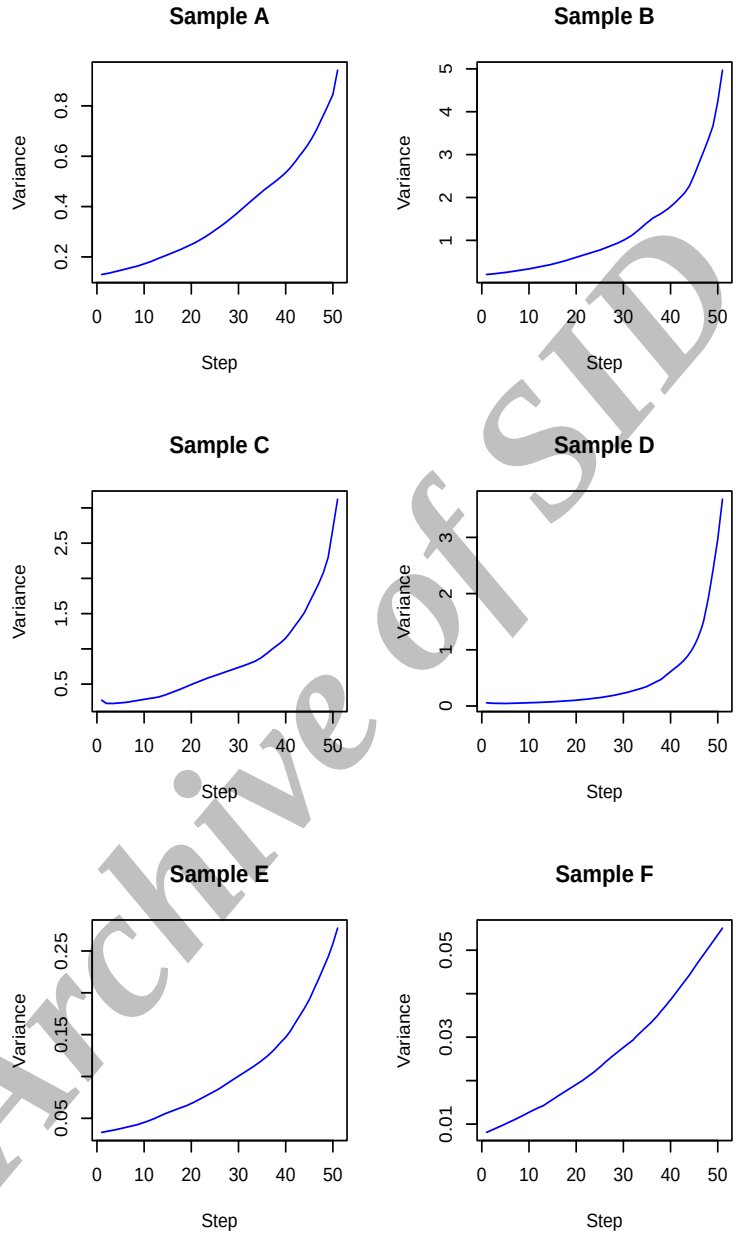
برای بررسی اثر مشاهدات پرت نمونه‌هایی در نظر گرفته می‌شوند که شامل مشاهده پرت باشند، برای این کار شش نمونه قبل به صورت زیر تغییر می‌یابند.

نمونه G : ۹۰ مشاهده از توزیع نرمال استاندارد و ۱۰ مشاهده از توزیع نرمال با پارامترهای $\mu = ۵$ و $\sigma = ۱$.

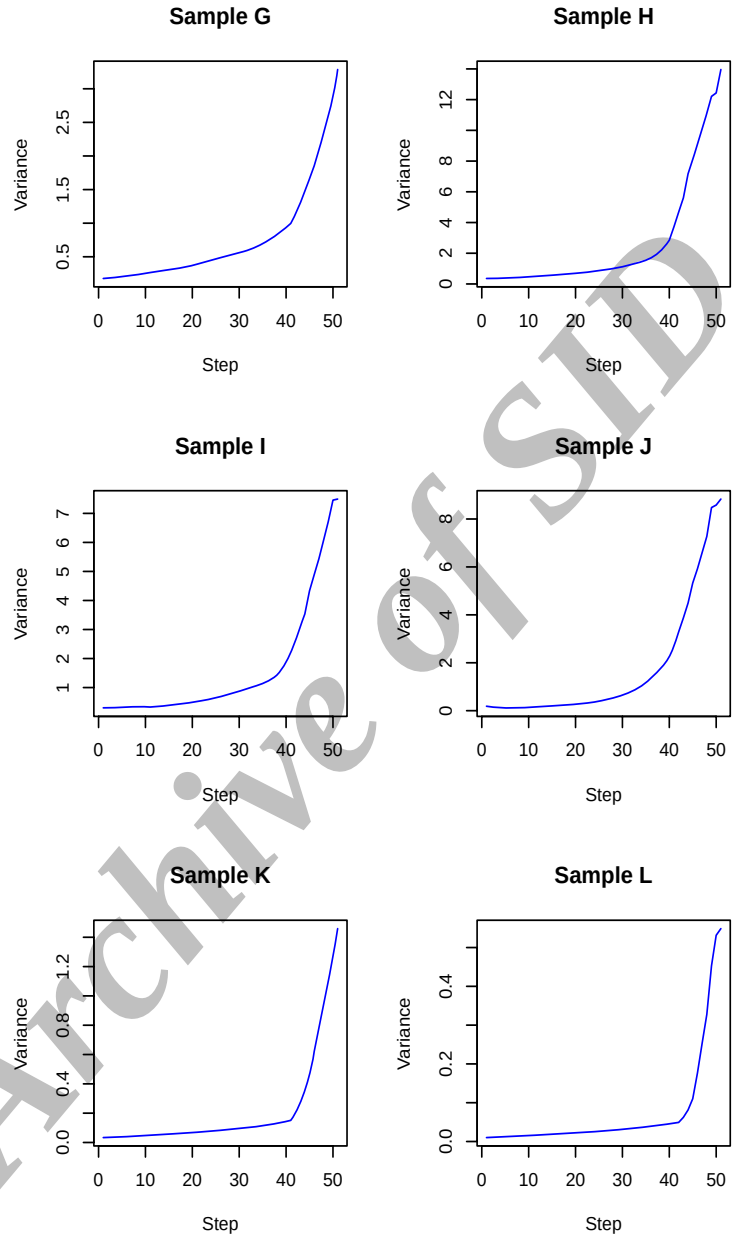
نمونه H : ۹۰ مشاهده از توزیع $t_{(df=۲)}$ و ۱۰ مشاهده از توزیع نرمال با پارامترهای $\mu = ۱۰$ و $\sigma = ۱$.

نمونه I : ۹۰ مشاهده از توزیع $\chi^{۲}_{(df=۲)}$ و ۱۰ مشاهده از توزیع نرمال با پارامترهای $\mu = ۱۰$ و $\sigma = ۱$.

نمونه J : ۹۰ مشاهده از توزیع $LogNorm(\mu = ۰, \sigma = ۱)$ و ۱۰ مشاهده از توزیع



شکل ۱: نمودار برآورد واریانس برای نمونه‌های فاقد مشاهده پرت در هر مرحله از جستجوی پیشرو



شکل ۲: نمودار برآورد واریانس برای نمونه‌های دارای مشاهده پرت در هر مرحله از جستجوی پیشرو

نرمال با پارامترهای $\mu = ۱۰$ و $\sigma = ۱$.

نمونه K : ۹۰ مشاهده از توزیع $Weibull(sh = ۲, sc = ۱)$ و ۱۰ مشاهده از توزیع نرمال با پارامترهای $\mu = ۵$ و $\sigma = ۱$.

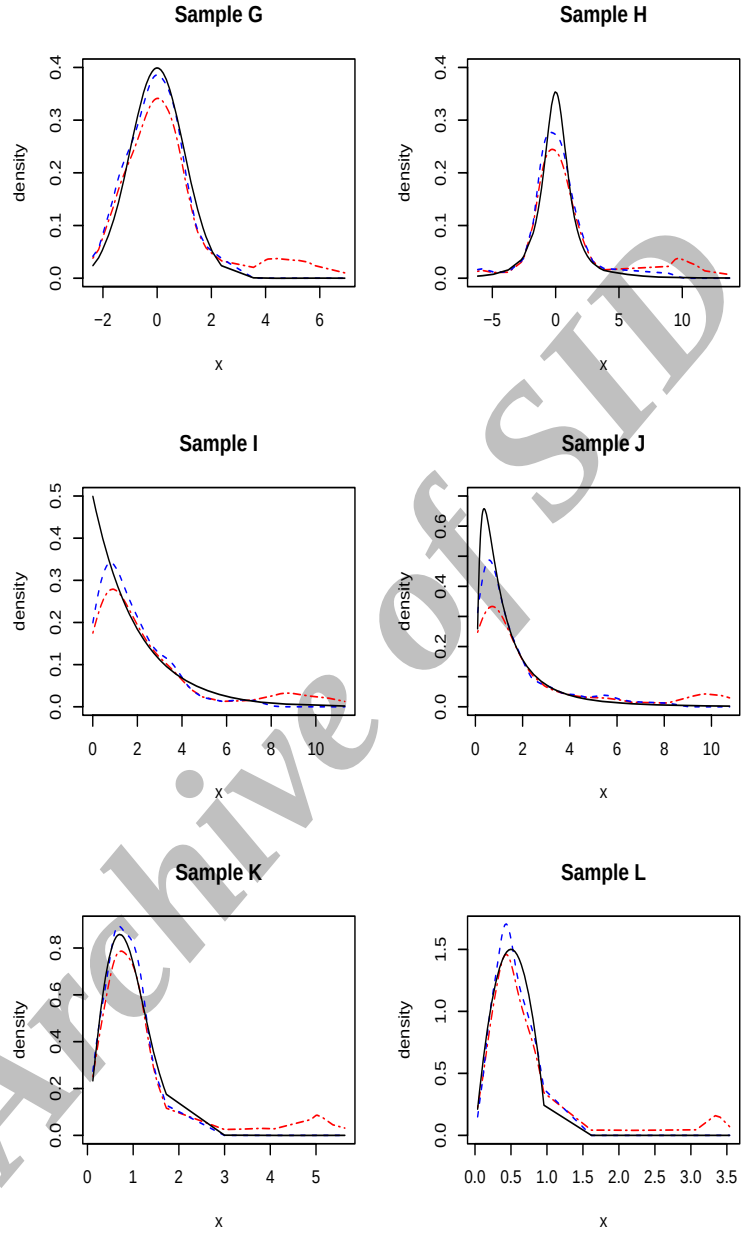
نمونه L : ۹۰ مشاهده از توزیع $Beta(\alpha = ۲, \beta = ۲)$ و ۱۰ مشاهده از توزیع نرمال با پارامترهای $\mu = ۲$ و $\sigma = ۱$.

شکل۲ نتایج جستجوی پیشرو را برای شش نمونه G تا L نشان می‌دهد. همانند شکل۱ محور افقی در این نمودارها، مراحل باقیمانده تا وارد شدن تمام مشاهدات و محور عمودی واریانس مشاهدات را در هر مرحله، نشان می‌دهند. همان‌طور که ملاحظه می‌شود از مرحله ۴۰ به بعد به دلیل وارد شدن مشاهدات پرت، برآورد واریانس به سرعت افزایش می‌یابد. البته این تغییر ناگهانی ممکن است به دلیل ساختار نمونه‌های مختلف، دقیقاً در مرحله ۴۰ اتفاق نیافتاده باشد. اما به هر حال نمودارهای ارائه شده در شکل۲ می‌تواند برای شناسایی داده‌های پرت به کار رود.

به منظور بررسی تاثیر مشاهدات پرت بر روی برآورد تابع چگالی، برای هر یک از نمونه‌های G تا L، نمودارهای برآورد تابع چگالی با در نظر گرفتن داده‌های پرت و با حذف داده‌های پرت در شکل ۳ رسم شده است. می‌توان این نمودارها را به ترتیب با نمودار تابع چگالی پیشنهادی برای اکثریت داده‌ها(یعنی تابع چگالی مربوط به نمونه انتخابی A تا F) مقایسه نمود. برای انجام مقایسه دقیقتر این دو برآورد از معیار میانگین توان دوم خطا[۵] (MSE) نیز استفاده شده است. نتایج این مقایسه‌ها، در جدول ۱ آورده شده است. در این جدول MSE1 نشان دهنده میانگین توان دوم خطای برآورد تابع چگالی با در نظر گرفتن مشاهدات پرت نسبت به تابع چگالی پیشنهادی برای اکثریت داده‌ها و MSE2 میانگین توان دوم خطای برآورد تابع چگالی با حذف داده‌های پرت نسبت به تابع چگالی پیشنهادی برای اکثریت داده‌ها می‌باشد. همان‌طور که جدول ۱ نشان می‌دهد، MSE2 برای تمام نمونه‌ها کمتر از MSE1 است و این مطلب بیانگر این است که حذف داده‌های پرت منجر به برآورد مناسب‌تری برای تابع چگالی می‌شود.

---

۵ Mean Square Error



شکل ۳: نمودار برآورد تابع چگالی در حضور مشاهدات پرت (نقطه و خط)، برآورد تابع چگالی با حذف مشاهدات پرت (خط چین) و تابع چگالی پیشنهادی اکثریت داده‌ها(خط ممتد) برای نمونه‌های مختلف

جدول ۱: میانگین مربع خطای برآورد تابع چگالی در حضور مشاهدات پرت (MSE1) و در غیاب مشاهدات پرت (MSE2) برای نمونه‌های مختلف

نمونه	MSE1	MSE2
G	۰/۰۰۱۸۷۷	۰/۰۰۰۴۳۵
H	۰/۰۰۲۸۸۴	۰/۰۰۱۶۶۷
I	۰/۰۱۱۳۱۶	۰/۰۰۷۷۰۰
J	۰/۰۲۹۵۲۲	۰/۰۰۷۷۳۵
K	۰/۰۰۳۷۹۱	۰/۰۰۲۳۳۵
L	۰/۰۲۷۷۵۲	۰/۰۲۱۴۵۶

بحث و نتیجه‌گیری

از آن‌جا که وجود مشاهدات پرت تاثیر زیادی در کارایی برآورد تابع چگالی دارد روشی برای شناسایی این مشاهدات لازم است. در این مقاله با استفاده از جستجوی پیشرو روشی برای تشخیص مشاهدات پرت ارائه شد. پس از شناسایی مشاهدات پرت، اگر تعداد نسبی آن‌ها زیاد نباشد، با حذف این مشاهدات از داده‌ها، می‌توان برآورد مناسب‌تری برای تابع چگالی به دست آورد. برآورد حاصله، با حذف داده‌های پرت، دارای کارایی بالاتری خواهد بود.

تقدیر و تشکر

نویسندگان از داوران محترم که پیشنهادات ارزنده ایشان موجب بهبود مقاله گردید، کمال تشکر و سپاسگزاری را دارند.

مراجع

Atkinson, A. C. (1994), Fast Very Robust Methods for the Detection of Multiple Outliers, *Journal of the American Statistical Association*, **89**,

1329-1339.

Atkinson, A. C. and Riani, M. (2000), *Robust Diagnostic Regression Analysis*, Springer-Verlag, New York.

Atkinson, A. C., Riani, M. and Cerioli, A. (2010), The Forward Search: Theory and Data Analysis, *Journal of the Korean Statistical Society*, **39**, 117-134.

Bertaccini, B. and Varriale, R. (2007), Robust Analysis of Variance: An Approch Based on the Forward Search, *Computational Statistics and Data Analysis*, **51**, 5172-5183.

Coin, D. (2008), Testing Normality in the Presence of Outliers, *Statistical Methods and Applications*, **17**, 3-12.

Hadi, A. S. (1992), Identifying Multiple Outliers in Multivariate Data, *Journal of the Royal Statistical Society*, **54**, 761-771.

Hadi, A. S. and Simonoff, J. S. (1993), Procedures for the Identification of Multiple Outliers in Linear Models, *Journal of the American Statistical Association*, **88**, 1264-1272.

Parzen. (1962), On Estimation of Probability Density Function and Mode, *Annals of Mathematical Statistics*, **3**, 1065-1076.

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

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Journal of Statistical Sciences, Autumn & Winter 2009 **224**

Vol. 3, No. 2, pp 199-209.

Estimation of Density Function in the Presence of Outliers

Mahdavi, A. and Towhidi, M.

Department of Statistics, Shiraz University, Shiraz, Iran.

Abstract: One of the most important issues in inferential statistics is the existence of outlier observations. Since these observations have a great influence on fitted model and its related inferences, it is necessary to find a method for specifying the effect of outlier observations. The aim of this article is to investigate the effect of outlier observations on kernel density function estimation. In this article we have tried to represent a method for identification of outlier observations and their effect on kernel density function estimation by using forward search method.

Keywords: Kernel Density Function Estimation, Forward Search Method, Smoothing Parameter.

Mathematics Subject Classification (2000): 62G07, 62G05