فصل ۱۴

14-1 - روشهای تخمین تابع توزیع

14-2 - روش χ^2-Test

14-3 - مدل مارکوف و χ^2-Test

14-4 - Kolmogorov-Smirnov Test

14-5 - فاصله Kullback-leibler دو تابع توزیع

14-6 - Anderson-Darling Test

14-1 روشهای تخمین تابع توزیع

ضرورت این کار عمده برای مقایسه و تخمین تابع توزیع متغیرهایی است که از نتایج آزمایش و شبیه سازی بدست آمده وجود دارد

14-1-a روش Naive Estimator

این روش به هیستوگرام معروف است. برای محاسبه تابع توزیع در این روش ابتدا مبدا را x_0 گرفته و به صورت bin (bin) به عرض h در نظر می گیریم که به صورت بر سطح ها به صورت $[x_0+mh, x_0+(m+1)h]$ برای m های مثبت و منفی تعریف می گردد.

فرض کنید تابع توزیع واقعی داده ها $\phi(x)$ باشد. در مسائلی که در این

فصل ارائه می‌شود تخمینی از $\phi(x)$ را خواهند داد که با $\hat{\phi}(x)$ نشان

می‌دهیم. در روش Naive تابع توزیع $\phi(x)$ با $\hat{\phi}(x)$ به صورت

زیر تخمین زده می‌شود.

$$\hat{\phi}(x) = \frac{1}{nh}\left(\text{تعداد } x_i \text{ های در یک bin که } x \text{ نیز در آن قرار دارد}\right)$$

این روش حساس به این است که مرزهای تابع باز یا بسته انتخاب شوند. اگر سایز h هر bin را کاهش دهیم $\hat{\phi}(x)$ به یک سری شکستگی های زیاد تبدیل می‌شود. حتی ممکن است هیچ داده‌ای در bin‌ای که سایز کوچک دارد قرار نگیرد. اگر h را نیز بزرگ بگیریم مثلاً همه داده ها در یک bin قرار گیرند تابع توزیع به یک خط تبدیل می‌شود. انتخاب اندازه h کار ساده‌ای نیست.

به جز این‌که این روش حساس به باز و بسته بودن توابع bin است تخمین $\hat{\phi}(x)$ حساس به انتخاب h می‌باشد.

Kernel Estimator 14-1-b

روش بعدی در تخمین، کرنل معروف است. در این روش $\phi(x)$ با $\hat{\phi}(x)$

که به صورت زیر تعریف می‌شود تخمین زده می‌شود

$$\hat{\phi}(x) = \frac{1}{nh} \sum_{i=-\frac{n}{2}}^{n/2} k\left(\frac{x-x_i}{h}\right)$$

$k(x)$ معمولاً تابع متقارن حول x می‌باشد و کرنل نام دارد. برای اینکه

$\hat{\phi}$ نرمالیزه باشد ($\int \hat{\phi}(x)dx = 1$) نتیجه می‌دهد که

$$\int_{-\infty}^{+\infty} k(x)\,dx = 1$$

کرنل استانداردی که استفاده می‌شود گاوسی می‌باشد یعنی: $k(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$.

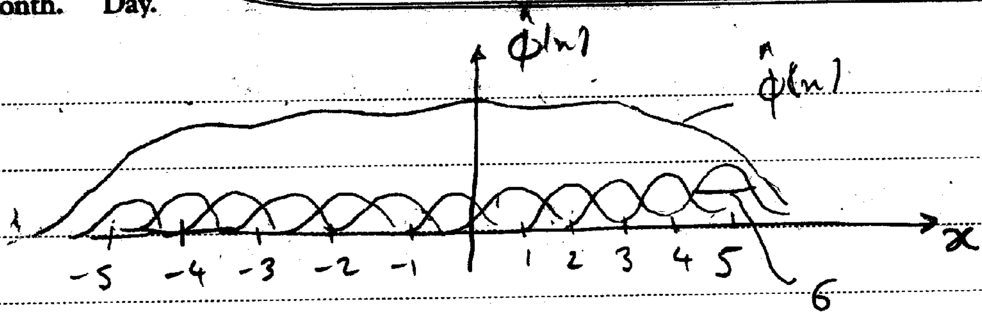
در واقع در پیوستگی $k(x)$، $\hat{\phi}(x)$ نیز پیوسته خواهد بود و کرنل اثر

هر نمونه را به نواحی i-ام منتقل می‌کند. برای کرنل گاوسی این اثر

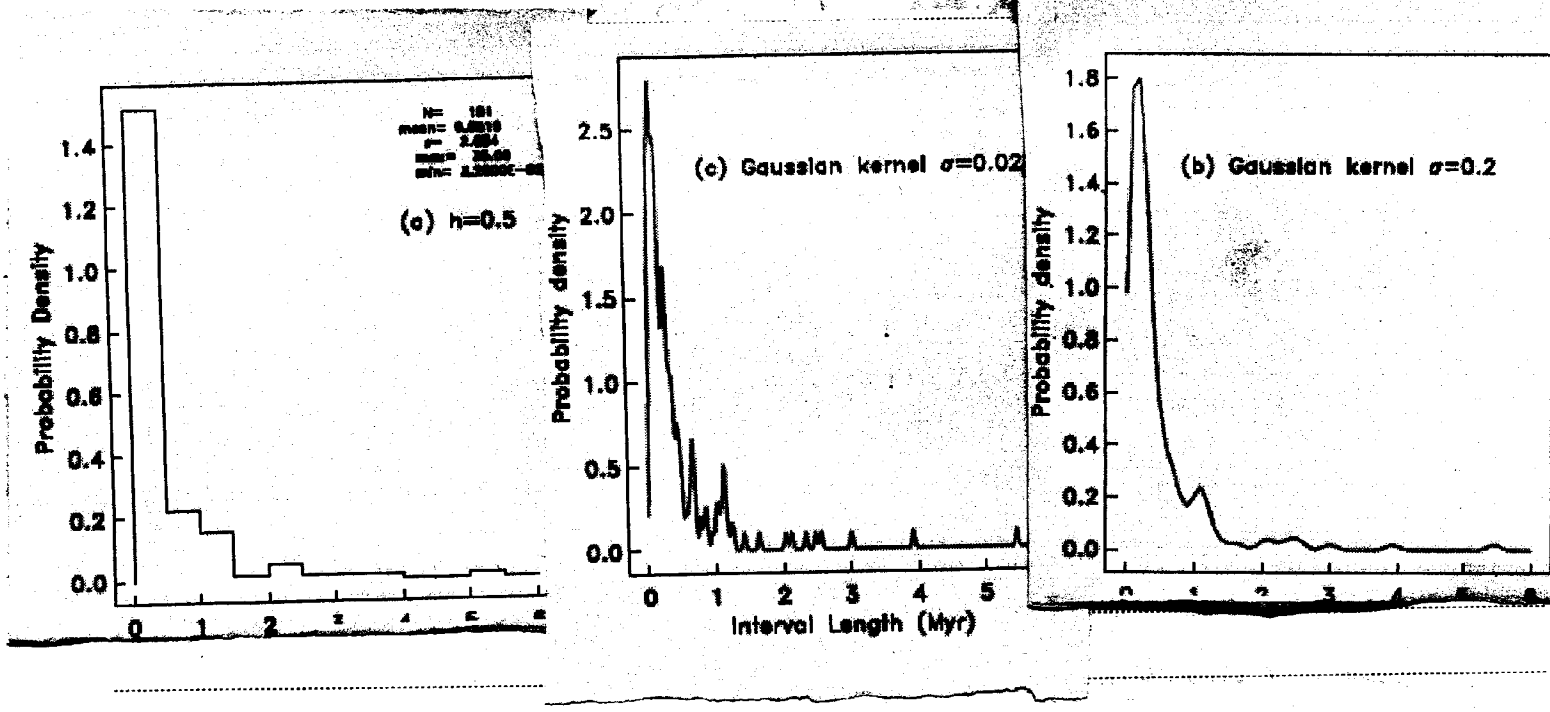
با $e^{-(x-x_i)^2}$ کم خواهد شد. هر چه σ (که در اینجا همان h است)

بزرگتر، نرم کننده‌ی بهتری خواهد بود. شکل نوعی اثر کرنل

در شکل صفحه بعد آمده که اثر هر نمونه در هر bin واضح است.



که در این روش نقش سایز bin در روش Naive را دارد. گرچه این روش
کاربرد فراوانی دارد ولی شکل $\hat{\phi}(x)$ تابع σ می‌شود که هیچ کنترلی
روی σ و مقدار مناسب آن نداریم. در شکل‌های زیر تخمین تابع توزیع
به روش Naive، $h=0.5$ و روش kernel با $\sigma=0.2$ ،
$\sigma=0.02$ نشان داده شده است که واضح است با افزایش σ، $\hat{\phi}(x)$
نرم‌تر خواهد شد.



معمولاً با تکرار ؛ افزایش n هم که در این دو روش از دو تابع توزیع بیشتر مطلع شد ولی در عوض این کار از دست دادن دقت در مرکز توزیع خواهد بود که داده زیادی داریم. دو روش دیگر روش ارائه شده است:

انتخاب اندازه bin به ~~[illegible]~~ [illegible] از

Sample distribution function 14-1-C

در این روش در عمل ؛ مجموعه T_n که به صورت زیر است

$$T_n = \{x_1, - , x_n\}$$

موجود است در اختیار داریم. تابع توزیع تجمعی $S_n(x)$ را به صورت زیر تعریف می‌کنیم

$$S_n(x) = \frac{1}{n}\left[\# \text{ of } x_i \in T_n < x\right]$$

این توزیع را وقتی بدست می‌آوریم که x_i ها را از نزولی به صعودی مرتب کنیم. در واقع $S_n(x_{(i)})$ به صورت خواهد بود داریم

$$S_n(x_{(i)}) = \frac{i}{n}$$

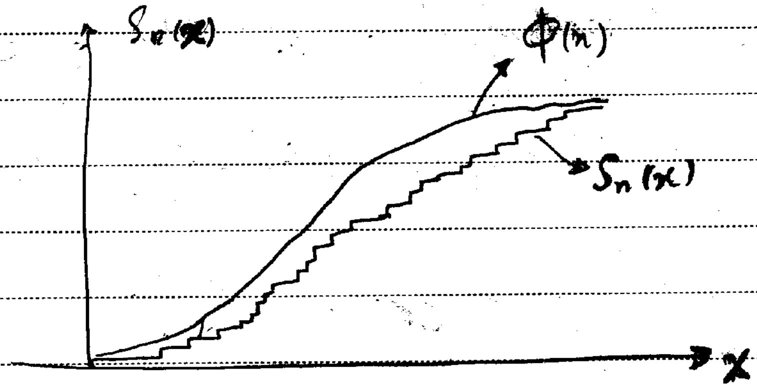
که n کل داده هاست، همچنین تابع بر x هایی که در داده‌ها اتفاق می‌افتند دارای پرش به اندازه $\frac{1}{n}$ خواهد بود. دقت کنید $S_n(x)$

خواهد کامل تابع توزیع را دارد. اگر n به اعداد بزرگ میل کند انتظار می‌رود که $S_n(x)$ تابع پیوسته شده و به $\Phi(x)$ میل کند.

این را قانون اعداد بزرگ می‌نامیم که

$$\lim_{n \to \infty} S_n(x) \longrightarrow \Phi(x)$$

در شکل زیر $S_n(x)$ نوعی رسم شده است.

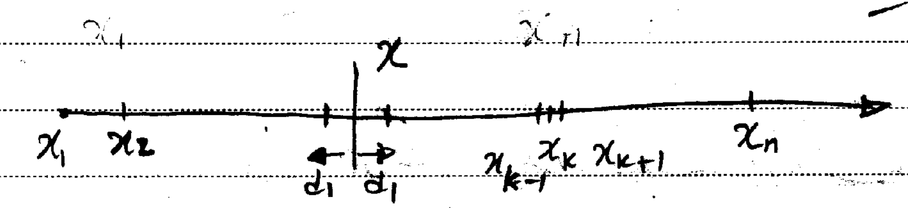


گاهی $\Phi(x)$ را از یک تئوری می‌دانیم و $S_n(x)$ را می‌توانیم به‌جای آن مقایسه کنیم. در واقع فاصله $\Phi(x)$ و $S_n(x)$ را کم کنیم. معیارهای متداول برای محاسبه فاصله بین دو تابع معرفی می‌گردد که χ^2 رتبه kalmogorov-Simirnov و فاصله kullback-leibler می‌باشد.

Nearest Neighbors Method. 14-1-d

معمولاً دقت می‌رسیم به کل ناحیه x که binها را بدست آوردیم در عمل یکسان نخواهد بود. در واقع در وسط توزیع دقت بالا و در دم‌ها دقت کمتری داریم. روش‌های نزدیک و کرنل تغییر که در بخش بعد معرفی خواهد شد سعی دارند این ایراد را رفع کنند.

در روش نزدیک‌ترین همسایه ایده اصلی این است که سعی کنیم تعداد داده‌های هر bin یکسان باشد. یعنی در وسط تابع توزیع hها را کوچک و در دم‌ها بزرگ بگیریم. در عمل ابتدا داده‌ها را از کوچک به بزرگ مرتب می‌کنیم.



حال نقطه x را نسبت به این محور در نظر می‌گیریم. برای هر نقطه x ما حالا می‌توانیم فاصله x و x_i داده‌ها را به صورت زیر تعریف کنیم

$$d_i = |x_i - x|$$

طوری که $d_1(x) \le d_2(x) \le \dots \le d_n(x)$ باشد.

حال $\hat{\phi}(x)$ که تخمین تابع توزیع k همسایه نزدیک می‌باشد به صورت زیر تعریف می‌شود:

$$\hat{\phi}(x) = \frac{(k-1)}{2n\,d_k(x)}$$

که برای اولین همسایه $(k=2)$ خواهد بود.

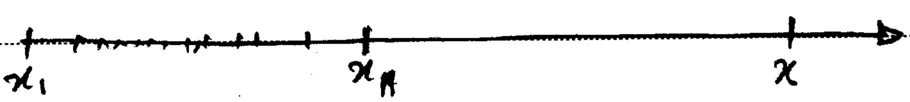
برای دیدن دلیل این تعریف دقت می‌کنیم که تعداد مشاهدات که در ناحیه $[x-r, x+r]$ به $r>0$ برابر خواهد بود با $2rn\phi(x)$.

بنابراین در ناحیه $[x-d_k(x), x+d_k(x)]$ ما به تعداد $(k-1)$ مشاهده خواهیم داشت. بنابراین $(k-1) = 2d_k(x)\phi(x)$ بدست خواهد آمد.

دقت کنید در نزدیکی‌ها d_k کوچک است و در دم‌ها بزرگ خواهد بود.

مثل روش Naive تابع توزیع $\hat{\phi}(x)$ نرمی نخواهیم داشت.

ایراد اصلی این روش اینست که وقتی x در ناحیه بیرون (x_1, x_n) قرار بگیرد. مثلاً در شکل زیر x بزرگتر از x_n می‌باشد.



بنابراین $d_k(x)=|x-x_n|$ (فاصله اولین همسایه) در صورتی که x به خود x میل می‌کند و بنابراین اگر $d(x) \to x$ خواهد بود و $\hat{f}(x) \sim \frac{1}{x}$ می‌شود. تابع توزیعی که در بینهایت به $\frac{1}{x}$ میل کند نرمالیزه شده نخواهد بود.

14-1-E (روش کرنل متغیر)

در این روش $d_{i,k}$ را به صورت فاصله نقطه x_i از k-امین نقطه نزدیک ($k \geq 1$) در مجموعه داده‌ها تعریف می‌شود

$$d_{i,k} = x_{i+k} - x_i$$

$\hat{f}(x)$ از رابطه زیر تخمین زده می‌شود.

$$\hat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{n d_{i,k}} K\left(\frac{x-x_i}{h d_{i,k}}\right)$$

در واقع اگر نقاط با فاصله زیادی از یکدیگر باشند (دم تابع توزیع) $d_{i,k}$ بزرگ است و در وسط تابع توزیع $d_{i,k}$ کوچک‌تر می‌شود و پنجره کرنل را کنترل می‌کند.

The χ^2 – Test 14-2

یکی از روشهایی که برای اندازه‌گیری فاصله بین توزیع اندازه‌گیری شده و توزیع خاص مورد نظر χ^2 است. بسته به نوع مسئله می‌توان χ^2 های مختلفی را تعریف کرد که در زیر چند مثال آورده شده است.

به عنوان اولین مثال فرض کنید که اندازه‌گیری یک کمیت مقدار O_i را بدست دهد و انتظاری که مانند اندازه‌گیری داشته باشیم (احتمالا از یک تئوری ناشی شود) E_i باشد. برای بررسی نزدیکی O_i به E_i χ^2 زیر را تعریف می‌کنیم

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

که k تعداد bin ها خواهد بود. اگر O_i و E_i تابع توزیع باشند می‌توان O_i و E_i را تابع توزیع تجمعی در نظر گرفت که در بخش 14-1-C گفته شد و χ^2 فاصله دو تابع توزیع تجمعی را خواهد داد.

به عنوان مثال دیگر فرض کنید O_i و E_i تعداد داده‌ها در bin i-ام باشد. انتظار ما از تئوری که توزیع $P(x)$ دارد این تعداد به صورت خواهد بود از $NP(x_j)$. فرض کنید تعداد اندازه‌گیری شده در bin i-ام، $h(x_j)$ باشد پس χ^2 مربع‌ها را می‌توانیم به صورت زیر نوشت کنیم:

$$\chi^2 = \sum_{j=1}^{n} \frac{\left[h(x_j) - NP(x_j)\right]^2}{\sigma_j^2(h)}$$

$\sigma_j(h)$ انحراف معیار در bin i-ام است. اگر یک آرایه داشته باشیم می‌توان $\sigma_j(h)$ را محاسبه کنیم. راه ساده که $\sigma_j(h)$ را بدست آوریم از توزیع دوجمله‌ای است. یعنی اگر یک عدد رندم را از آرایه برداریم و سؤال کنیم که آیا این داده متعلق به bin i-ام است یا نه؟ احتمال وقوع این حالت از توزیع دوجمله‌ای بدست می‌آید. یعنی عدد انتخاب شده در bin i-ام است یا نه. احتمال اینکه در خانه i-ام باشد $P(x_j)$ خواهد بود و احتمال نبودن $q(x_j) = 1 - P(x_j)$ خواهد بود. انحراف معیار عبارت است از $\sqrt{NPq}$ که با توجه به کم بودن P برای تابع عبارت است از $\sqrt{NP}$ که

Subject:

Year. 200 Month. Day.

که فقط همان $Np(x_j)$ بجای $h(x_j)$ باشد. بنابراین

$$\chi^2 = \sum_{j=1}^{n} \frac{[h(x_j) - Np(x_j)]^2}{Np(x_j)}$$

$$\simeq \sum_{j=1}^{n} \frac{[h(x_j) - Np(x_j)]^2}{h(x_j)}$$

اگر $h(x_j) = Np(x_j)$ باشد، بنابراین $\chi^2 = 0$ خواهد بود و در عمل این کار امکان‌پذیر نیست که در همه binها این ارتباط برقرار باشد.

بنابراین χ^2 عدد مثبت خواهد بود. وقتی یک آرایه را که دارای min و Max است به بازه‌هایی با ابعاد خاص و به تعداد n تقسیم می‌کنیم می‌توانیم در هر bin $h(x_j)$ را اندازه بگیریم. در بعضی از binها ممکن است داده‌ای وجود نداشته باشد. تعداد binهایی که در آن داده وجود دارد و $h(x_j)$ وجود دارد درجات آزادی نام دارد که با ν نمایش می‌دهیم. اگر χ^2_ν را به صورت $\chi^2_\nu = \frac{\chi^2}{\nu}$ تعریف کنیم مرتبه χ^2_ν از $O(1)$ خواهد بود.

14-3 تخمین طول مارکف از روش χ^2

اگر یک اندازه گیری داشته باشیم و بخواهیم چک کنیم که آیا آرایه بدست آمده دیتای چه طول مارکفی است می توانیم از روش χ^2 استفاده کنیم. اگر تابع توزیع سه نقطه ای Joint را بنویسیم خواهیم داشت:

$$P_3(x_3,t_3;x_2,t_2;x_1,t_1)=P_{1|2}(x_3,t_3|x_2,t_2;x_1,t_1)\times P_2(x_2,t_2;x_1,t_1)$$

اگر پروسه مارکف با طول ۱ داشته باشیم

$$P_{3\,\text{Markov}}(x_3,t_3;x_2,t_2;x_1,t_1)=P_{1|1}(x_3,t_3|x_2,t_2)\times P_2(x_2,t_2;x_1,t_1).$$

حال می توانیم χ^2_ν را به صورت زیر تعریف کنیم

$$\chi^2_\nu=\frac{1}{\nu}\int\frac{\left[N P_3(x_3,t_3;x_2,t_2;x_1,t_1)-N P_{3\,\text{Markov}}(x_3,t_3;x_2,t_2;x_1,t_1)\right]^2}{N P_3(x_3,t_3;x_2,t_2;x_1,t_1)}\times dx_1\,dx_2\,dx_3$$

که فرض $NP(x_j)$ همان $h(x_j)$ می‌باشد. بنابراین

$$\chi^2 = \sum_{j=1}^{n} \frac{[h(x_j) - NP(x_j)]^2}{NP(x_j)}$$

$$\simeq \sum_{j=1}^{n} \frac{[h(x_j) - NP(x_j)]^2}{h(x_j)}$$

اگر $h(x_j) = NP(x_j)$ باشد بنابراین $\chi^2 = 0$ خواهد بود. در عمل این کار امکان پذیر نیست که در همه binها این ارتباط برقرار باشد. بنابراین χ^2 عدد مثبت خواهد بود. وقتی یک آزمایش را که دارای min و Max است به بازه‌های با ابعاد خاص و به تعداد n تقسیم می‌کنیم می‌توانیم در هر bin $h(x_j)$ را اندازه بگیریم. در بعضی از binها شاید داده‌ای وجود نداشته باشد. تعداد binهایی که در آن داده وجود دارد و $h(x_j)$ وجود دارد درجات آزادی نام دارد که با ν نمایش می‌دهیم. اگر χ_ν^2 را به صورت $\chi_\nu^2 = \frac{\chi^2}{\nu}$ تعریف کنیم مرتبه χ_ν^2 از $O(1)$ خواهد بود.

برای بررسی مانا می‌توانیم ابتدا t_1 را صفر و $t_2 = \frac{t_3}{2}$ بگیریم (یعنی $t_3 = 2t_2$).

سپس χ^2_ν فقط تابعی از t_2 خواهد بود که Min χ^2_ν حول

زمان مارکف را خواهد داد. در زیر دو منحنی χ^2_ν برای

CMB و return قیمت نفت آورده شده است که در

CMB به جای t_2 زاویه Θ داریم. در Min χ^2 برای مارکف

است یعنی زاویه حدود 1° به Θ_{Markov}. در شکل دوم برای قیمت نفت

χ^2_ν از ابتدا افت و خیز می‌کند و Min مطلق ندارد که نشان

می‌دهد حول زمان 1 (یعنی داده‌ها) افت و خیز کرده

اگر x_i قیمت نفت باشد $y_i = \ln \frac{x_{i+1}}{x_i}$ را return [illegible]

در بازار آمریکا.

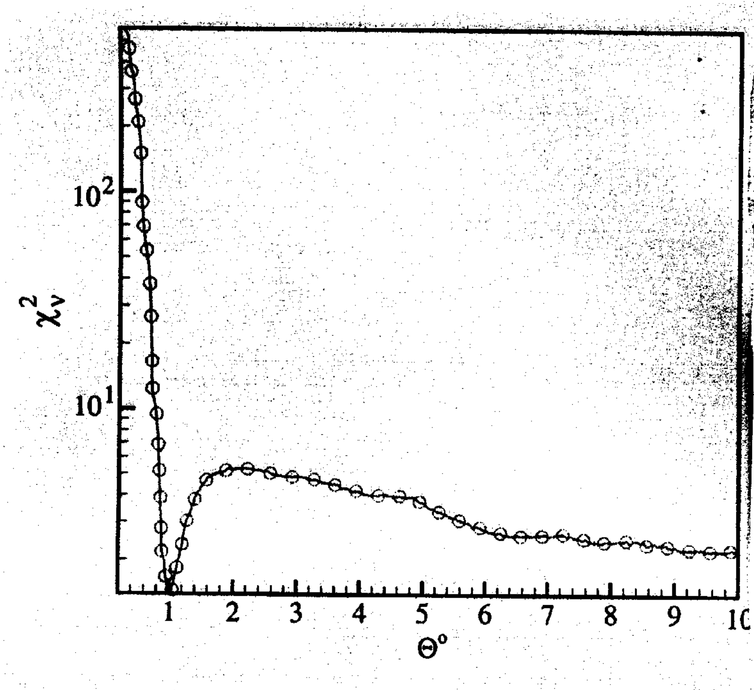


Figure 3. χ^2_ν in terms of the angular scale Θ. The minimum value of $\chi^2_\nu = 1.388$ corresponds to $\Theta_{\mathrm{Markov}} = 1.01^{+0.09}_{-0.07}$, with a 1σ confidence level.

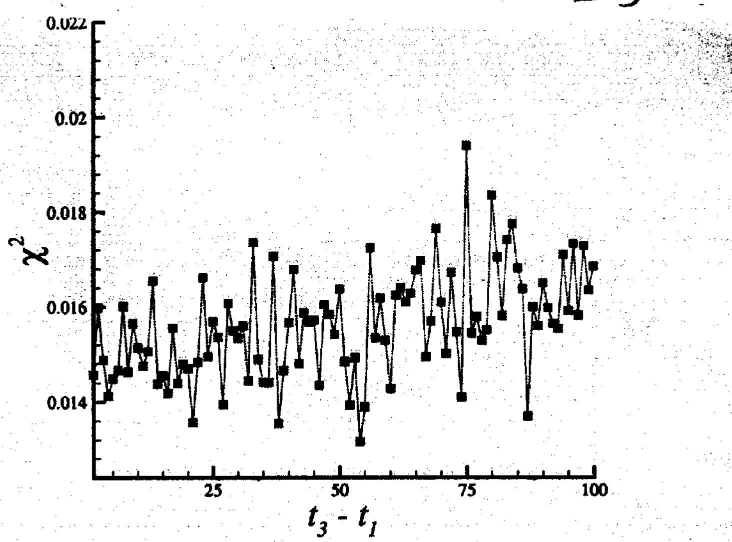


FIG. 3. Test of the Chapman-Kolmogorov equation for the time series of the returns, indicating that the Markov time scale t_M is 1.

به عنوان آخرین مثال فرض کنید می‌خواهیم چک کنیم آیا دو آرایه وابسته توزیع هستند یا نه؟

فرض کنید تعداد داده‌ها در bin ام برای خانواده اول $h(x_j)$ و برای خانواده دوم $g(x_j)$ باشد. χ^2 را به صورت زیر تعریف می‌کنیم

$$\chi^2 = \sum_{j=1}^{n} \frac{[g(x_j) - h(x_j)]^2}{\sigma^2(g) + \sigma^2(h)}$$

که $\sigma^2(g)+\sigma^2(h)$ واریانس تفاضل $g(x_j)-h(x_j)$ خواهد بود.

Kolmogorov-Smirnov Test 14-4
(KS-Test)

می‌توانیم در بخش 14-1-c دیدیم نیاز به تعریف فاصله بین ۲ توزیع تجربی و تئوری داریم. اولین و ساده‌ترین روش همان χ^2 می‌باشد که در بخش قبل دیدیم. روش بعدی KS-Test است که به صورت زیر تعریف می‌شود. فرض کنید آرایه $x_1, \ldots, x_N$ که مرتب شده (از کمتر به بیشتر) باشند. S_N را به صورت زیر تعریف می‌کنیم

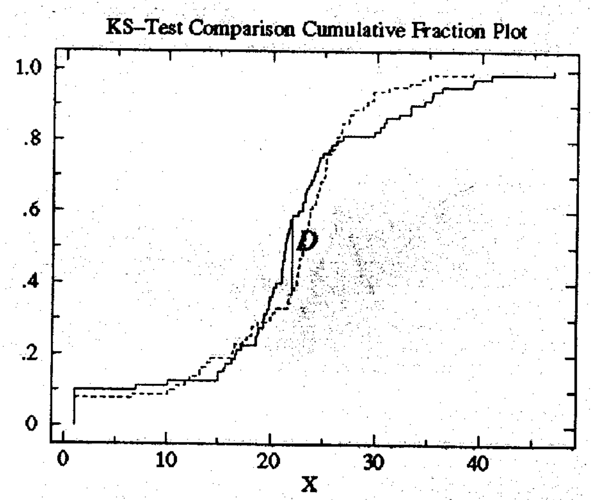
$$S_N = \frac{n(i)}{N}$$

که $n(i)$ تعداد نقاط کمتر از x_i می‌باشد. که $n(i)=i$ می‌باشد.

اگر S_n را بخواهیم با یک تئوری که دارای تابع توزیع تجمعی F باشد مقایسه کنیم فاصله D دو تابع توزیع را به صورت زیر تعریف می‌کنیم.

$$D = \max_{1 \leq i \leq N} \left(F(x_i) - \frac{i-1}{N} , \frac{i}{N} - F(Y_i) \right)$$

فاصله D در شکل زیر نشان داده شده است.



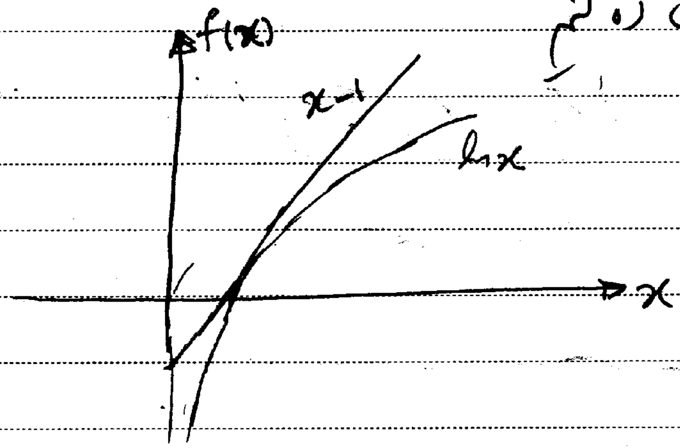
فاصله kullback-leibler 14-5 (KL)

فاصله KL یا آنتروپی KL به صورت زیر بین دو تابع توزیع مثلاً تئوری $P(x_i)$ و تجربی $q(x_i)$ به صورت زیر تعریف می‌شود.

$$d_{pq} = -\sum_{j=1}^{n} p(x_j) \ln \frac{q(x_j)}{p(x_j)}$$

دقت می‌کنیم که این فاصله تقارن ندارد یعنی $d_{pq} \neq d_{qp}$ (در حالت کلی).

و d_{pq} همیشه مثبت می‌باشد. برای دیدن این خاصیت از نامساوی $\ln x \leq x-1$ استفاده می‌کنیم



$$D = \sum_{j=1}^{n} p(x_j) \ln \frac{q(x_j)}{p(x_j)} \leq \sum p(x_j) \left(\frac{q(x_j)}{p(x_j)} - 1 \right)$$

$$= \sum q(x_j) - \sum p(x_j) = 1 - 1 = 0$$

پس $D \leq 0$

و لذا $d_{pq} \geq 0$

که $d_{pq} = -D$ می‌باشد. برای مثال می‌توان طول

مارکوف را از این توصیف به دست آورد یعنی

$$d = -\int dx_1 dx_2 dx_3 \; P_3(x_3,t_3;x_2,t_2;x_1,t_1) \times \ln\left(\frac{P_{3Markov}(x_3,t_3;x_2,t_2;x_1,t_1)}{P_3(x_3,t_3;x_2,t_2;x_1,t_1)}\right)$$

که بجز $t_1=0$ ، $t_3=2t_2$ خواهد بود که Min ، d

$t_2=t_M$ ، زمان مارکف را خواهد داد.

Anderson-Darling Test تست -۱۴-۶
(AD-Test)

این تست روش برای مقایسه تابع توزیع تجمعی F در پیشتر S_N که تعمیم از تست KS می باشد. در واقع مشخص شده با توزیع گذاریم F تغییرات را کم کرده و بدین وسیله سهم بیشتر به دم خواهد داد. این تست با رابطه زیر داده می شود.

$$A^2 = -N - S$$

که

$$S = \sum_{j=1}^{N} \frac{(2j-1)}{N} \left[\ln F(X_j) + \ln \left(1 - F(X_{N+1-j})\right) \right]$$

که F تابع توزیع تجمعی است، (داده‌ها X_i ها مرتب شده‌اند).

مقادیر بحرانی A برای دو توزیع معروف F در جداول خاصی موجود

هستند که می‌توانید به سایت زیر مراجعه کنید

http://www.itl.nist.gov/