بسم الله الرحمن الرحيم

# همگرایی ضعیف فرایندهای تجربی و آزمون نیکویی برازش برای مدلهای پارامتری

توسط

مریم قدسی

رساله ارائه شده به عنوان بخشی از ملزومات برای دریافت درجه

کارشناسی ارشد آمار

دکتر محمد قاسم وحیدی اصل

استاد مشاور

دکتر خلیل شفیعی



دانشکده علوم ریاضی

دانشگاه شهید بهشتی

شهریورماه ۱۳۷۹

بسمه تعالی

تاریخ ..................

شماره ..................

پیوست ..................

دانشگاه شهید بهشتی

دانشکده علوم ریاضی

## صورتجلسه دفاع از پایان نامه

جلسه هیأت داوران ارزیابی پایان نامه خانم مریم قدسی

به شماره شناسنامه : ۶۳۳ صادره‌از : شیراز متولد : ۱۳۵۳ دانشجوی دوره

کارشناسی ارشد ناپیوسته رشته : آمار

با عنوان : همگرایی فرآیند تجربی پارامتری تعمیم یافته و آزمونهای نیکویی برازش برای مدلهای پارامتری

به راهنمائی : آقای دکتر محمدقاسم وحیدی‌اصل طبق دعوت قبلی در تاریخ ۷۹/۶/۱۳

تشکیل گردید و براساس رأی هیأت داوری و با عنایت به ماده ۲۰ آئین نامه کارشناسی

ارشد مورخ ۷۳/۱۰/۲۵ پایان نامه مزبور با نمره ۱۹ (نوزده تمام) و درجه عالـــی

مورد تصویب قرار گرفت .

| | |
|---|---|
| ۱- آقای دکتر محمدقاسم وحیدی‌اصل | استاد راهنما |
| ۲- آقای دکتر خلیل شفیعی | استاد مشاور |
| ۳- آقای دکتر علی عمیدی | استاد داور |
| ٤- آقای دکتر احمد خدادادی | استاد داور |
| ۵- | |

# قدردانی

در این‌جا لازم است از کلیه‌ی افرادی که مرا در انجام این پایان‌نامه کمک نموده‌اند، خصوصاً از اساتید گرامی جناب آقای دکتر محمد قاسم وحیدی اصل و سرکار خانم دکتر زهرهٔ شیشه‌بر که در تمام مراحل انجام این پایان‌نامه با مساعدت‌ها و راهنمایی‌های بی‌دریغ خود مرا یاری کردند، تشکر کنم.

# همگرایی ضعیف فرایندهای تجربی و آزمون نیکویی برازش برای مدلهای پارامتری


## چکیده

می‌خواهیم توزیع یک سری دادهٔ خام را مشخص نماییم. برای این منظور از آزمونهای معنی‌داری سود برده، ابتدا آمارهٔ آزمون و سپس توزیع آن را تحت مدل فرض شده به‌دست می‌آوریم. از همگرایی ضعیف فرایندهای تجربی استفاده کرده و برای دو نوع داده، داده‌های بدون سانسور (کامل) و سانسور شده، آزمون را انجام می‌دهیم.

برای داده‌های کامل یا بدون سانسور، آمارهٔ کولموگرف ـاسمیرنف بهترین آماره می‌باشد.

برای داده‌های زمان شکست، برآورد کاپلان ـمایر تعمیم‌یافتهٔ $F_n(t)$ را با تابع توزیع برآورد شدهٔ $F(t, \hat{\theta})$ تحت مدل فرض شده، که $\hat{\theta}$ برآورد ماکزیمم درستنمایی است، مقایسه می‌کنیم. فرایند تجربی پارامتری تعمیم‌یافتهٔ $\sqrt{n}(F_n(t) - F(t, \hat{\theta}))$ توسط انتگرالهای تصادفی نسبت به یک مارتینگل تقریب زده می‌شود. این در توزیع به یک فرایند گاوسی همگراست که می‌تواند، توسط انتگرالهای، تصادفی نسبت به یک حرکت براونی نمایش داده شود. همگرایی ضعیف فرایند تجربی پارامتری روی خط کلی برای داده‌های سانسور شدهٔ تصادفی به‌دست آورده می‌شود. از طریق تبدیل فرایند گاوسی حدی به یک مارتینگل گاوسی، آزمونهای خوبی برازندگی را برای مدلهای پارامتری به دست می‌آوریم.



واژه‌های کلیدی: آزمون کولموگرف ـاسمیرنف، برآورد کاپلان ـمایر تعمیم یافته، تبدیل، تقریب مارتینگل، حرکت براونی، دادهٔ سانسور شده.

# فهرست مندرجات

# لیست اشکال

فصل ۱

# مقدمه

برای برآورد تابع توزیع مجموعه‌ای از داده‌ها، روشهای پارامتری و ناپارامتری متفاوتی به کار می‌روند. روشهای ناپارامتری تعداد محدودی بیش نیستند که در زیر آنها را می آوریم.
فرض کنید $X_1, X_2, X_3, \ldots$ متغیرهای تصادفی و هم‌توزیع (i.i.d)باشند به طوری که

$$X \cong (\mu, \sigma^2)$$

یعنی توزیع $X$، دارای میانگین $\mu$ و واریانس $\sigma^2$ است. همچنین فرض کنید

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

قانون قوی اعداد بزرگ $(SLLN)$ می‌گوید که وقتی $n \to \infty$

$$\bar{X}_n \xrightarrow{a.s} \mu$$

(این رابطه بدون هیچ گونه فرضی در مورد واریانس برقرار است) پس هرگاه که $n$ زیاد می‌شود و سیر صعودی پیدا می‌کند میانگین را از روی داده‌ها، با دقت، می‌توان برآورد کرد.
ارزیابی دقت این تقریب توسط قضیهٔ حد مرکزی داده‌می‌شود به طوری که وقتی $n \to \infty$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1) \qquad (1-1)$$

سرعت همگرایی در (۱—۱) وقتی $E|X|^{3} < \infty$ به صورت

$$\sup_x |P\Big(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\Big) - \Phi(x)| \leq \frac{33}{4}\frac{E|X-\mu|^{3}}{\sigma^{3}}\frac{1}{\sqrt{n}}$$

تعیین می‌شود که از نتیجهٔ بری- اسن به‌دست می‌آید و $\Phi$ تابع توزیع نرمال استاندارد می‌باشد.

یکی از مفیدترین نتایج نظریهٔ احتمال قضیهٔ من- والد است که برای هر تابع پیوسته $g$ و با استفاده از (۱—۱) هنگامی که $n$ به سمت بینهایت می‌رود، خواهیم داشت:

$$g\Big(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\Big)\xrightarrow{d} g\Big(N(0, 1)\Big)$$

به عنوان مثال وقتی $n \to \infty$،

$$n\Big(\frac{\bar{X}_n - \mu}{\sigma}\Big)^{2} \xrightarrow{d} \chi^{2}(1)$$

که البته این نتیجه کلی می‌باشد یعنی اگر

$$X_n \xrightarrow{d} X$$

آنگاه برای هر تابع پیوستهٔ $g$ داریم:

$$g(X_n) \xrightarrow{d} g(X)$$

قضایای دیگری نیز وجود دارند که سرعت همگرایی $\bar{X}_n - \mu$ را به صفر ارائه می‌دهند، یکی از آنها قانون لگاریتم مکرر (LIL) است که وقتی $n \to \infty$ می‌گوید:

$$\sqrt{n}(\bar{X}_n - \mu)/b_n \to [-\sigma, \sigma] \quad a.s.$$

که در آن $b_n = \sqrt{2\log(\log n)}$ و $\to$ بدین معناست که برای تقریباً هر $\omega$ ($a.e.\ \omega$) در مجموعهٔ $\Omega$ از فضای احتمال اصلی $(\Omega, \mathcal{F}, P)$ که متغیرهای تصادفی روی آن تعریف شده‌اند، دنبالهٔ اعداد حقیقی $\sqrt{n}(\bar{X}_n(\omega) - \mu)/b_n$ دارای نقاط حدی بازهٔ $[-\sigma, \sigma]$ می‌باشند.[۴۲]

قانون لگاریتم مکرر به صورت زیر داده می‌شود:

$$\underline{\lim}_{n\to\infty} \max_{1\leq k\leq n} \frac{1}{\sqrt{n}}|\sum_{i=1}^{k}(X_i - \mu)|b_n = \frac{\pi}{2} \quad a.s.$$

اگر $X_1, X_2, \ldots$ متغیرهای تصادفی $i.i.d$ دلخواهی باشند، وقتی $n \to \infty$ داریم:

$$n^{-1} \log P(\bar{X}_n \geq 0) \to \log \rho$$

به طوری که

$$\rho = \inf_t M(t)$$

که $M(t)$ تابع مولد گشتاوری است ($M(t) = E(e^{tX})$)و بیشتر این که اگر $X$، در شرطهای استاندارد

$$t^* = \sup\{t : M(t) < \infty\} \leq \infty \quad , \quad M'(0^+) < \infty$$

و برای $0 < t < t^*$

$$M'(t) > 0$$

صدق کند، نتیجهٔ قوی‌تر زیر به دست خواهیم آورد:

$$P(\bar{X}_n \leq 0) = \frac{\rho^n b_n}{\sqrt{n}}$$

که در آن $0 < \underline{\lim} b_n \leq \overline{\lim} b_n < \infty$.[۸]

مطالعهٔ تابع توزیع تجربی با روشهای ناپارامتری فوق وسعت داده می‌شود، اما این روشهای ناپارامتری در بعضی از عملکردها کارا نیستند و تعداد این روشها به تعداد محدود معرفی شده در بالا خلاصه می شوند و شرایط بسیاری از مسائل با این روشها جور نمی‌باشند، بنابراین از روش جامع‌تر تابع توزیع تجربی استفاده می‌کنند. که قضیهٔ گلیونکو- کانتلی [۱] نیکویی این برآورد را تصدیق می‌کند.

در بررسی داده‌های زمان شکست (مثلاً طول عمر یک لامپ) نیز توزیع فراوانی زمانهای بقا معمولاً از توزیع نرمال حتی با انجام تبدیل ساده‌ای هم فاصله دارد و تابع توزیع تجربی در این حالت، برآورد کاپلان- مایر می‌باشد که مستقیماً از روی داده‌ها به دست می آید و در بررسی الگوهای آماری مفید می‌باشد.

در این پایان‌نامه ابتدا فرایندهای تجربی را در حالت کلی تعریف کرده و فرایند تجربی پارامتری تعمیم‌یافته را که برای داده‌های زمان شکست است، ارائه می‌دهیم. که برای این منظور بایستی تا حدودی به مفاهیم آنالیز بقا آشنا باشیم بنابراین در بخش دوم فصل سوم به آن می‌پردازیم.

---

[۱] Glivanko-Cantelli

اغلب اوقات در حد، صحبت از خود متغیر امکان‌پذیر نیست، اما می‌توان از توزیع آن در حد صحبت کرد، برای این منظور از همگرایی ضعیف استفاده می‌نمایند. در فصل چهارم ابتدا همگرایی ضعیف فرایند تجربی به یک پل براونی و سپس همگرایی ضعیف فرایند تجربی پارامتری تعمیم‌یافته به یک فرایند گاوسی را بیان می‌کنیم که این فرایند گاوسی را می‌توان توسط انتگرالهای تصادفی نسبت به یک فرایند وینر نمایش داد.

در پایان از طریق تبدیل این فرایند گاوسی حدی به یک مارتینگل گاوسی، رده‌ای از آزمونهای نیکویی برازش برای مدلهای پارامتری را ارائه می‌دهیم.

لازم به تذکر است که برای مطالعهٔ این دفتر بایستی با مفاهیم اولیهٔ فرایندهای تصادفی آشنا بود که ابتدای فصل دوم را به این بحث اختصاص داده‌ایم. همچنین همگرایی ضعیف فرایندهای تصادفی و انتگرال تصادفی ابزار کار ما در این جا می‌باشند و تا جایی که رفع ابهام نمایند در بخش سوم و چهارم از فصل دوم توضیح داده می‌شوند.

همچنین برهان و اثبات قضایا و لم‌هایی که در طول فصلها وجود دارند و طولانی می‌باشند، در پایان همان فصل مربوطه آورده می‌شوند.

فصل ۲

# تعریف‌ها و پیش‌نیازها

## ۱.۲ عنصر تصادفی و فرایندهای تصادفی خاص

فرض کنید $M$ مجموعه‌ای از تابعهای $x$ باشد که هر $t$ در مجموعهٔ $T$ را به عدد حقیقی $x(t)$ مرتبط می‌کند. $T$ معمولاً یک مجموعهٔ اقلیدسی مانند $[۰,۱]$ و یا $R$ است.

برای هر عدد صحیح $k \geq ۱$ و هر $t_۱, \ldots, t_k$ در $T$، فرض کنید $\pi_t = \pi_{t_۱,\ldots,t_k}$ نگاشت $M$ به توی فضای $k$ـ بعدی $R^k$ باشد که به صورت زیر تعریف می‌شود:

$$\pi_t(x) = \pi_{t_۱,\ldots,t_k}(x) = (x(t_۱), \ldots, x(t_k)). \qquad (۲-۱)$$

آنگاه برای هر زیر مجموعهٔ بورل در کلاس $B^k$ از زیر مجموعه‌های بورل $R^k$. مجموعهٔ $\pi^{-۱}_{t_۱,\ldots,t_k}(B)$ یک «زیر مجموعهٔ متناهی بعدی $M$» نامیده می‌شود.

حال $\mathcal{M}_۰$ که شامل همهٔ زیر مجموعه‌های متناهی بعدی $M$ می‌باشد لزوماً یک میدان است، چون در واقع می‌توان $\pi^{-۱}_{t_۱,\ldots,t_k}(B)$ را به صورت

$$\pi^{-۱}_{t_۱,\ldots,t_k}(B) = \{x \in M \ : \ (x(t_۱), \ldots, x(t_k)) \in B\}$$

نوشت.

فرض کنید $\mathcal{M}$ سیگما میدان تولید شده توسط میدان $\mathcal{M}_۰$ از زیرمجموعه‌های متناهی بعدی باشد. برای چنین مجموعهٔ $M$، فضای اندازه‌پذیر $(M, \mathcal{M})$ را فضای تابعی اندازه‌پذیر

روی $T$ می‌نامیم. یک نگاشت اندازه‌پذیر از فضای احتمال $(\Omega, \mathcal{F}, P)$ به یک فضای تابعی اندازه‌پذیر (یا به یک فضای اندازه‌پذیر دلخواه) را عنصر تصادفی روی (یعنی مجموعهٔ مقادیر خود را در) این فضای اندازه‌پذیر می‌نامند.

تعریف (۲–۱). عنصرهای تصادفی روی فضاهای تابعی اندازه‌پذیر را همچنین یک فرایند تصادفی می‌نامند.

مسئله (۲–۱). فرض کنید $\mathcal{B} = \mathcal{B}(R)$ زیر مجموعه‌های بورل خط حقیقی $R$ باشد. آنگاه $X$ یک فرایند تصادفی روی $(M, \mathcal{M})$ است، اگر و فقط اگر برای هر $\omega \in \Omega$، $X(., \omega) \in M$ و برای هر $t \in T$، $X(t, .)$ یک متغیر تصادفی روی $(R, \mathcal{B})$ باشد.($X$ یک متغیر تصادفی روی $(R, \mathcal{B})$ نامیده می‌شود اگر و فقط اگر برای هر $B \in \mathcal{B}$، $X^{-1}(B) \in \mathcal{F}$.)

برهان. توجه می‌کنیم که $X$ بایستی یک نگاشت اندازه‌پذیر از $(\Omega, \mathcal{F}, P)$ به $(M, \mathcal{M})$ باشد، به عبارت دیگر بایستی نگاشت

$$(t, \omega) \xrightarrow{X} X(t, \omega)$$

اندازه‌پذیر باشد یعنی

$$X^{-1}(B) \in \mathcal{B} \times \mathcal{F}$$

که از قضیهٔ فوبینی نتیجهٔ مورد نظر حاصل می‌گردد. □

این عنصرتصادفی اندازهٔ احتمال $P_X$ را روی جفت $(M, \mathcal{M})$ توسط رابطهٔ

$$P_X(A) = P(X \in A) = P(X^{-1}(A)) \quad ; A \in \mathcal{M} \qquad (۲–۲)$$

تولید می‌کند(برای زیر مجموعه‌های متناهی بعدی $A$).

دو متغیر تصادفی $X$ و $Z$ را «هم‌ارز» می‌نامند اگر دارای توزیعهای یکسان باشند.

دو فرایند $X$ و $Z$ دارای توزیعهای متناهی بعدی یکسان هستند اگر $P_X$ و $P_Z$ روی $\mathcal{M}_\circ$ منطبق باشند. وقتی که $X$ و $Z$ دارای توزیعهای متناهی بعدی یکسان باشند، آنها را فرایندهای «هم‌ارز» می‌نامند.