بسم الله الرحمن الرحیم

دانشگاه شهید بهشتی

دانشکده علوم ریاضی

پایان‌نامه دوره کارشناسی ارشد آمار ریاضی

روش‌های ناپارامتری در تحلیل سری‌های زمانی ناخطی

نگارنده

پروین جلیلی

استاد راهنما

دکتر مجتبی خزائی

استاد مشاور

دکتر محمدرضا فرید روحانی

مرداد ۱۳۸۸

تقدیم به

نور جاویدان زمین و زمان،

خانوادهٔ عزیزم

و همه کسانی که دوستشان دارم.

قدردانی و تشکر

این همه گفتیم لیک اندر بسیچ — بی‌عنایات خدا هیچیم، هیچ

بی‌عنایات حق و خاصان حق — گر ملک باشد، سیاهش ورق

ای خدا! ای فضل تو حاجت روا — با تو یاد هیچ‌کس نبود روا

اینقدر ارشاد، تو بخشیده‌ای — تا بدین، بس عیب ما پوشیده‌ای

قطره‌ای دانش که بخشیدی ز پیش — متصل گردان به دریاهای خویش

قطره‌ای علم است اندر جان من — وارهانش از هوا وز خاک تن

(مولانا)

از استاد ارجمندم جناب آقای دکتر مجتبی خزائی که زحمت راهنمایی این رساله را پذیرفته و بدون شک انجام مراحل مختلف این رساله بدون راهنمایی‌های گرانبها و تلاش‌های ارزندهٔ ایشان ممکن نبود، کمال تشکر و قدردانی را دارم.

از استاد گرامی جناب آقای دکتر محمدرضا فریدروحانی که مشاورهٔ این رساله را بر عهده داشتند صمیمانه تشکر و قدردانی می‌نمایم.

از اساتید محترم جناب آقای دکتر مجتبی گنجعلی و جناب آقای دکتر رضا حبیبی که زحمت مطالعه و داوری این رساله را تقبل فرمودند کمال تشکر را دارم.

از تمامی دوستان عزیزم که مرا در انجام این رساله همراهی نمودند صمیمانه کمال تشکر را دارم.

گر سر هر موی من یابد زبان — شکرهای تو نیاید بر زبان

چکیده

در این رساله به مطالعهٔ مدل‌ها و روش‌های ناپارامتری در تحلیل سری‌های زمانی پرداخته شده است. در سال‌های اخیر، به دلیل انعطاف‌پذیری مدل‌ها و روش‌های ناپارامتری در مدل‌بندی و تحلیل سری‌های زمانی، تمایل به استفاده از این مدل‌ها در مقایسه با مدل‌های پارامتری افزایش یافته است. در این رساله ضمن مروری کلی بر روش‌های ناپارامتری در تحلیل سری‌های زمانی، مدل اتورگرسیو با ضرایب تابعی (FAR) با جزئیات بیشتری بررسی شده است. روش رگرسیون خطی موضعی برای برازش این مدل و روش میانگین خطای پیش‌بینی برای انتخاب پارامتر هموارسازی استفاده شده است. از سوی دیگر، آزمون‌هایی برای بررسی خوبی برازش مدل (FAR) در مقابل مدل‌های پارامتری و ناپارامتری معرفی شده‌اند. یک روش پیش‌بینی خودگردان برای پیش‌بینی نقطه‌ای و فاصله‌ای با استفاده از روش رگرسیون خطی موضعی برای مدل اتورگرسیو با ضرایب تابعی پیشنهاد شده است. در انتها با تشریح دو مثال واقعی، کارایی مدل معرفی‌شده در مقابل مدل‌های پارامتری خطی و ناخطی بررسی شده است. نتایج به‌دست آمده، گویای عملکرد بهتر مدل ارائه‌شده در مقابل مدل‌های پارامتری است. این مدل برای برازش به داده‌های سری زمانی قیمت سبد نفتی ایران استفاده و پیش‌بینی‌های نقطه‌ای و فاصله‌ای ارائه شده است.

کلید واژه‌ها: روش رگرسیون ناپارامتری، هموارساز، مسئلهٔ بلای بعد زیاد، مدل جمعی، سری‌های زمانی ناخطی، مدل اتورگرسیو با ضرایب تابعی.

پیشگفتار

مشاهدات سری‌های زمانی از جمله مشاهداتی هستند که تقریباً در تمامی حوزه‌ها وجود داشته و تحلیل و بررسی آنها یکی از زمینه‌های مهم در علم آمار و برخی علوم دیگر است. شاید بتوان گفت مهمترین هدف در مطالعه سری‌های زمانی کمک به کشف مکانیزم حاکم بر سیستمی است که منجر به تولید مشاهدات می‌شود. در راستای نیل به این هدف از سالیان دراز تلاش‌های بسیاری صورت گرفته، به خصوص در تولد و تکامل مدل‌های سری‌های زمانی خطی گاوسی می‌توان به یول (۱۹۲۷) در مدل‌بندی اتورگرسیو (AR) مشاهدات لکه‌های خورشیدی، تا کار باکس و جنکینز (۱۹۷۰) که بطور برجسته‌ای مدل‌بندی ARMA را هم از بعد نظری و هم از نظر روش‌شناسی تکامل بخشیدند، اشاره کرد. در هر حال مدل‌بندی رفتارهای غیراستاندارد دیگر مانند غیرنرمال بودن، دوره‌های غیرمتقارن، روابط ناخطی بین متغیرهای تاخیری، خارج از محدوده مدل‌های سری‌های زمانی خطی است. چنین رفتارهایی در بسیاری از مشاهدات واقعی دیده می‌شوند. از جمله مشاهداتی که در این زمینه بسیار معروف هستند می‌توان به مشاهدات لکه‌های خورشیدی و مشاهدات گربه‌های سیاه‌گوش کانادایی (تانگ، ۱۹۹۰) اشاره کرد. خارج از محدوده روابط خطی تنوع روابط ناخطی بسیار زیاد است. پیشرفت‌های اولیه تحلیل سری‌های زمانی ناخطی روی بعضی روابط ناخطی

پارامتری تمرکز یافت. در میان مدل‌های ارائه شده می‌توان به مدل‌های موفقی چون مدل دوخطی (گرنجر و اندرسون، ۱۹۷۸)، مدل اتورگرسیو نمائی (EXPAR) (هاگن و اوزاکی، ۱۹۸۱ و اوزاکی، ۱۹۸۲) و مدل‌بندی ARCH (انگل، ۱۹۸۲ و بلرسلف، ۱۹۸۶) و مدل آستانه‌ای (تانگ، ۱۹۹۰، تیائو و تیسی، ۱۹۹۴) اشاره کرد.

در تحلیل سری‌های زمانی ناخطی وقتی که پیشینه‌ای از ساختار مشاهدات در دست نباشد شناسایی یک مدل ناخطی مناسب برای برازش به مشاهدات با توجه به گستردگی چنین روابط ناخطی می‌تواند بسیار سخت باشد. از طرفی با افزایش توان محاسباتی در سال‌های اخیر تحلیل مشاهدات سری‌های زمانی با اندازه‌های بسیار بزرگ و پیچیدگی‌های زیاد، قابل اجرا شده است. در چنین شرایطی اعتبار و کارایی مدل‌بندی پارامتری سوال برانگیز است. یک جایگزین مناسب برای مدل‌های ناخطی پارامتری در مواجهه با چنین شرایطی استفاده از مدل‌ها و روش‌های تحلیل ناپارامتری است. این روش‌ها علاوه بر اینکه شیوهٔ مستقلی برای تحلیل سری‌های زمانی فراهم می‌کنند، می‌توانند به درک و تشخیص بهتری از یک مدل پارامتری مناسب نیز منجر شوند. همچنین در مواجهه با سری‌های دراز مدت با پیچیدگی‌های زیاد، توانایی تشخیص ساختار مناسب و پیش‌بینی با دقت قابل قبولی را دارند. البته استفاده از روش‌ها و مدل‌های ناپارامتری (کامل)، زمانی که تعداد متغیرهای تاخیری و یا متغیرهای کمکی بیشتر از دو تا باشند نیاز به مشاهدات بسیار زیادی دارند، این مشکل تحت عنوان مسئلهٔ بلای بعد زیاد توسط بل‌من (۱۹۶۱) مطرح شد. برای غلبه بر این مشکل محدودیت‌هایی را روی مدل ناپارامتری (کامل) اعمال می‌کنند که این سبب به‌وجود آمدن مدل‌های مفیدی مانند مدل جمعی (هاستی و تیبشیرانی، ۱۹۹۰) و مدل رگرسیون با ضرایب تابعی (چن و تیسی، ۱۹۹۳ و کای و همکاران، ۲۰۰۰) شده است. این راهکارها ضمن غلبه بر مشکل

بالای بعد زیاد، انعطاف‌پذیری قابل قبولی در برازش به مشاهدات واقعی دارند. در این رساله ضمن مروری کلی بر روش‌ها و مدل‌های ناپارامتری در تحلیل سری‌های زمانی، مدل اتورگرسیو با ضرایب تابعی با جزئیات بیشتر شرح داده شده است.

در فصل اول این رساله، با توجه به شباهت زیاد روش‌های ناپارامتری در تحلیل سری‌های زمانی با روش‌های ناپارامتری که در تحلیل رگرسیونی استفاده می‌شود مروری اجمالی بر مدل‌ها و روش‌های ناپارامتری در تحلیل رگرسیون ساده و رگرسیون چندگانه خواهیم داشت. برای این منظور با روش هموارسازی نمودار پراکنش و چند هموارساز نمودار پراکنش مانند هموارساز هسته، هموارساز رگرسیون وزنی موضعی و اسپلاین رگرسیونی برای تحلیل رگرسیون ساده و مدل جمعی و مدل رگرسیون با ضرایب تابعی و چگونگی برازش آنها برای تحلیل رگرسیون چندگانه آشنا می‌شویم.

در فصل دوم این رساله، با روش‌های ناپارامتری در تحلیل سری‌های زمانی آشنا می‌شویم. این روش‌ها در سه حالت تحلیل سری‌های زمانی وقتی Y_t در مقابل زمان t مدل می‌شود، تحلیل سری‌های زمانی وقتی Y_t به یک متغیر تاخیری (Y_{t-d}) وابسته باشد و تحلیل سری‌های زمانی وقتی که Y_t به بیش از یک متغیر تاخیری وابسته است، مورد بررسی قرار می‌گیرند. در حالتی که Y_t به یک متغیر وابسته باشد روش هموارسازی هسته با جزئیات بیشتر و زمانی که Y_t به بیش از یک متغیر وابسته باشد مدل اتورگرسیو جمعی شرح داده می‌شود. همچنین در این فصل با اصل سفیدسازی از طریق پنجره‌بندی، فرایندهای آمیخته و روش میانگین خطای پیش‌بینی برای انتخاب پارامترهای مدل آشنا می‌شویم.

در فصل سوم این رساله، با مدل رگرسیون با ضرایب تابعی در تحلیل سری‌های زمانی و به خصوص با مدل اتورگرسیو با ضرایب تابعی آشنا می‌شویم. در این فصل پس از بحث در

شرایط مانایی و ارگودیک بودن فرایند حاصل از مدل اتورگرسیو با ضرایب تابعی، روش رگرسیون خطی موضعی برای برازش مدل، روش میانگین خطای پیش‌بینی برای انتخاب پارامترهای مدل، آزمون‌هایی برای بررسی کفایت برازش مدل اتورگرسیو با ضرایب تابعی در مقابل مدل‌های پارامتری و ناپارامتری و انتخاب بهترین مدل و روش‌هایی از جمله روش پیش‌بینی خودگردان برای پیش‌بینی سری زمانی برازش‌شده با این مدل ارائه می‌شوند.
در فصل چهارم این رساله نیز با ارائه سه مثال واقعی کارایی مدل و روش برازش معرفی‌شده در فصل سوم مورد بررسی قرار می‌گیرد. در مثال اول، با استفاده از داده‌های گربه‌های سیاه‌گوش کانادایی، توانایی مدل معرفی‌شده در مقابل مدل ناخطی پارامتری TAR بررسی می‌شود (کای و همکاران، ۲۰۰۰). در مثال دوم با استفاده از داده‌های میانگین ماهانه جریان حجمی آب در ایستگاه هیدرومتری پل دوآب—شرا از توابع استان مرکزی، توانایی مدل‌بندی مدل معرفی شده در مقابل مدل خطی پارامتری $SARIMA$ بررسی می‌شود و در نهایت از این مدل به منظور برازش مشاهدات سری زمانی قیمت سبد نفتی ایران استفاده شده و پیش‌بینی‌های نقطه‌ای و فاصله‌ای ارائه می‌شود.

فهرست مندرجات

فصل ۱

آشنایی با روش‌های ناپارامتری در تحلیل رگرسیونی

۱.۱ مقدمه

با توجه به شباهت زیاد روش‌های ناپارامتری در تحلیل سری‌های زمانی با روش‌های ناپارامتری که در تحلیل رگرسیونی استفاده می‌شود، در این فصل مرور اجمالی بر مدل‌ها و روش‌های ناپارامتری در تحلیل رگرسیون ساده و رگرسیون چندگانه خواهیم داشت. برای این منظور در بخش ۲.۱، به بررسی مدل و روش‌های ناپارامتری در تحلیل رگرسیون ساده خواهیم پرداخت. در این بخش با اصطلاح هموارسازی نمودار پراکنش و چند هموارساز نمودار پراکنش مانند هموارساز هسته، هموارساز رگرسیون وزنی موضعی و اسپلاین رگرسیونی آشنا می‌شویم.
در بخش ۳.۱ به بررسی مدل‌ها و روش‌های ناپارامتری در تحلیل رگرسیونی با چند متغیر پیشگو خواهیم پرداخت. در این بخش به معرفی مدل جمعی و مدل رگرسیون با ضرایب تابعی

و چگونگی برازش آنها می‌پردازیم.

۲.۱ رگرسیون ناپارامتری ساده

درحالتی که Y متغیر وابسته و X تنها متغیر پیشگو باشد مدل رگرسیون ناپارامتری ساده به صورت زیر نوشته می‌شود :

$$Y_i = m(X_i) + \epsilon_i \tag{۱.۱}$$

که در آن m تابعی نامعلوم و هموار (دارای مشتق پیوسته از مرتبه دوم) است. همچنین ϵ_iها متغیرهای تصادفی ناهمبسته با میانگین صفر و واریانس σ^2 هستند. و در مواردی $\epsilon_i \sim N(0, \sigma^2)$ فرض می‌شوند.

در مقایسه با مدل‌های رگرسیون پارامتری که شکل تابعی $m(X_i)$ مشخص و فقط شامل پارامترهای نامعلوم است در مدل (۱.۱) ویژگی‌های عمومی هموار بودن برای ضابطه $m(X_i)$ در نظر گرفته شده است. به همین دلیل مدل رگرسیونی ناپارامتری ساده انعطاف‌پذیری و قابلیت بیشتری در مدل کردن داده‌ها دارد. در تحلیل مدل‌های رگرسیون پارامتری برازاندن مدل، معادل برآورد پارامترهای آن است. این کار در مدل رگرسیونی ناپارامتری ساده معادل ترسیم یک منحنی هموار بر روی نمودار پراکنش Y در مقابل X است. بنابراین به برازاندن مدل رگرسیونی ناپارامتری ساده هموارسازی نمودار پراکنش نیز گفته می‌شود. از جمله هموارسازهایی که به منظور این هموارسازی استفاده می‌شود می‌توان به هموارساز میانگین گردشی، هموارساز هسته، هموارساز رگرسیون وزنی موضعی، اسپلاین رگرسیونی و اسپلاین مکعبی اشاره کرد.

تحلیل رگرسیون ناپارامتری ساده بر این اصل استوار است که مشاهدات متغیر پیشگو هر چه به نقطه $x_\circ$ نزدیک‌تر باشند شامل اطلاعات بیشتری از $m(x_\circ)$ هستند. بنابراین با ترکیب مشاهداتی که در همسایگی $x_\circ$ قرار دارند یا تخصیص وزن بیشتر به آنها می‌توان برآورد بهتری از $m(x_\circ)$ بدست آورد. از طرف دیگر در نگاهی موضعی و با توجه به ویژگی‌های ذکر شده برای $m(x)$ می‌توان $m(x_\circ)$ را در همسایگی $x_\circ$ توسط مقداری ثابت، یک پاره‌خط و در حالت کلی یک چندجمله‌ای از درجه پایین به خوبی تقریب کرد. اختلاف روش در هموارسازهای مختلف به نوع رفتار آنها به دو موضوع فوق (یعنی نوع وزن‌دهی و درجهٔ چندجمله‌ای برازش‌شده) برمی‌گردد. در ساده‌ترین حالت فرض می‌شود که $m(x)$ در همسایگی $x_\circ$ ثابت بوده و توسط مقداری ثابت تقریب می‌شود. مدل با این فرض، مدل ثابت موضعی نامیده می‌شود.

در صورتی که $W_i(x_\circ)$ وزن تخصیص داده شده به مشاهده iام باشد برآورد تابع رگرسیونی $m(x_\circ)$ به صورت میانگین وزنی مقادیر متغیر وابسته حول $x_\circ$ تعریف می‌شود و داریم :

$$\hat{m}(x_\circ) = \sum_{i=1}^{n} W_i(x_\circ) y_i \qquad (۲.۱)$$

توجه کنید که وزن iامین مشاهده یعنی $W_i(x_\circ)$ به مقدار $x_\circ$ بستگی دارد با انتخاب مناسب وزن‌ها این امکان فراهم می‌شود که مشاهدات نزدیک $x_\circ$ در برآورد $m(x_\circ)$ تاثیر بیشتری نسبت به مشاهدات دورتر داشته باشد. میانگین‌های وزنی یک انتخاب یا نتیجهٔ بسیاری از روش‌های مدل‌بندی ناپارامتری است و بطور کلی فرم $W_i(x_\circ)$ است که روش‌ها را از یکدیگر متمایز می‌نماید.

۱.۲.۱ هموارساز هسته

در این بخش با هموارساز هسته که در تحلیل سری‌های زمانی نیز کاربرد فراوان دارد آشنا می‌شویم. با توجه به مدل (۱.۱) تابع رگرسیونی $m(x)$ امید شرطی Y به شرط $X = x$ است و می‌توان نوشت

$$m(x) = E(Y|X = x).$$

اگر X و Y متغیرهای تصادفی پیوسته‌ای با چگالی $f(x, y)$ باشند داریم:

$$m(x) = E(Y|X = x) = \int_{-\infty}^{\infty} y f(y|x) dy = \int_{-\infty}^{\infty} y \frac{f(x, y)}{f(x)} dy, \qquad (۳.۱)$$

روشی برای برآورد $m(x)$، برآوردکردن چگالی‌های $f(x, y)$ و $f(x)$ و جایگزینی آنها در رابطهٔ فوق است. می‌دانیم که

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(x)dx = P(X \leq x) \\ f(x) &= \frac{dF(x)}{dx} = \lim_{h \to ۰+} \frac{F(x+h) - F(x-h)}{۲h}. \end{aligned} \qquad (۴.۱)$$

برآورد معمول $F(x)$ یعنی تابع توزیع تجربی به صورت زیر تعریف می شود.

$$F_n(x) = \frac{۱}{n} \sum_{i=۱}^{n} I(X_i \leq x)$$

که

$$I(X_i \leq x) = \begin{cases} ۱ & X_i \leq x \\ ۰ & \text{در غیر این صورت} \end{cases}$$

با جایگذاری $F_n(.)$ به جای $F(.)$ در (۴.۱) و برای یک $h > ۰$ مفروض، برآوردگر زیر برای $f_n(x)$ حاصل می‌شود.

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{۲h}$$

$$= \frac{۱}{۲nh}\sum_{i=۱}^{n}(I(X_i \leq x+h) - I(X_i \leq x-h)) \qquad (۵.۱)$$

با استدلی مشابه می‌توان برآوردگری نیز برای $f(x,y)$ یافت. تساوی (۵.۱) نشان می‌دهد برای یک h ثابت $f_n(x)$ تابعی پله‌ای از x است. به عبارت دیگر برآوردگر مناسبی برای تابع پیوسته و هموار $f(x)$ نیست. تساوی (۵.۱) را می‌توان به شکل زیر نوشت

$$f_n(x) = \frac{۱}{n}\sum_{i=۱}^{n}\frac{۱}{h}K(\frac{X_i - x}{h}) = \frac{۱}{n}\sum_{i=۱}^{n}K_h(X_i - x) \qquad (۶.۱)$$

که در آن $K(u) = \frac{۱}{۲}I(|u| \leq ۱)$ و $K_h(u) = \frac{۱}{h}K(\frac{u}{h})$. در این صورت $f_n(x)$ میانگین مقادیر تابع $K_h(u)$ در نقاط $(X_i - x)$ است. تابع $K(u)$ را یک هسته و چنین برآوردگری را برآوردگر هسته $f(x)$ با هسته $K(u)$ می‌نامند. در تساوی (۶.۱)، تابع هستهٔ $K(u)$ چگالی یکنواخت روی بازهٔ $(-۱,۱)$ است. در حالت کلی توابع هسته را چگالی‌های پیوسته و متقارن حول صفر تعریف می‌کنیم. چند تابع هسته معروف و پرکاربرد درجدول (۱.۱) آورده شده است.

$K(u)$	نام هسته
$(\frac{۳}{۴})(۱-u^۲)I(\|u\| \leq ۱)$	اپانچنیکوف
$(\frac{۱۵}{۱۶})(۱-u^۲)^۲I(\|u\| \leq ۱)$	مربعی
$(۱-\|u\|)I(\|u\| \leq ۱)$	مثلثی
$(\frac{۱}{۲\pi})exp(-\frac{u^۲}{۲})$	گاوسی
$(\frac{۱}{۲})I(\|u\| \leq ۱)$	یکنواخت

جدول ۱.۱: پرکاربردترین توابع هسته در مطالعات کاربردی

با ارائه تعریف مناسبی از تابع هسته دومتغیره می‌توان چگالی توام $f(x,y)$ را به صورت زیر برآورد کرد:

$$f_h(x,y) = \frac{۱}{n}\sum_{i=۱}^{n}K_h(X_i - x, Y_i - y),$$