مدل‌های خطی تابعی

سیدمحمدابراهیم حسینی‌نسب [۱]

چکیده:

تحلیل داده‌هایی با ماهیت تابعی که به فضاهای تابعی با بُعد نامتناهی متعلق هستند نیازمند بسترسازی مناسب و ارائه ابزارهای ویژه‌ای است. این کار از طریق آماده کردن نسخه‌های تصادفی از بعضی از نظریه‌های موجود در نظریه‌ی عملگرها یا آنالیز تابعی می‌باشد. هر چند که می‌توان این داده‌ها را با نگرشی چندمتغیره تحلیل نمود امّا به دلیل در نظر نگرفتن ماهیت تابعی داده‌ها این کار با اشکالاتی همراه است. در این مقاله ابتدا مفاهیم و تعاریف مربوط به مدل‌های خطی برای داده‌های تابعی را بیان و در ادامه نیز یک مجموعه داده‌ی واقعی (داده‌های دما و میزان بارندگی) تحلیل و نتایج آن گزارش می‌شود.

واژه‌های کلیدی: ویژه مقدار، ویژه تابع، تحلیل داده‌های تابعی، نظریه عملگرها، مدل خطی تابعی.

۱ مقدمه

در تحلیل داده‌های تابعی (FDA) داده‌ها به صورت تابعی از یک متغیر دیگر (معمولاً زمان) می‌باشند. به دلیل آنکه پیشرفتهای بوجود آمده در فن‌آوری باعث شده است تا مشاهده و جمع‌آوری این گونه داده‌ها امکان‌پذیر گردد، تحقیقات در این زمینه در چند سال اخیر با رشد فزاینده‌ای همراه بوده است. در این گرایش، قبلاً تحلیل و تلخیص اطلاعات یک نمونه از مشاهدات (توابع)، بیشتر به صورت توصیفی انجام می‌شده است، زیرا نظریه‌ها و استدلال‌های لازم برای استنباط‌های آماری، به اندازه‌ی نیاز تکامل پیدا نکرده بودند. امروزه هر چند که این نیاز کاملاً مرتفع نشده است اما تلاش‌هایی در این زمینه در حال انجام است. پذیره‌ی تعلق داده‌ها به فضاهای تابعی با بُعد نامتناهی، سبب می‌شود تا بحثها و نظریه‌های مربوطه در نظریه‌ی عملگرها یا آنالیز تابعی دنبال گردند. علاوه بر آن، این مسایل هنگامی که با پدیده‌ی تصادفی بودن توابع همراه شوند، پیچیدگی‌های موجود را دوچندان می‌کند.

شیمی سنجی، شاخه‌ای از علم شیمی که روشهای آماری را برای تحلیل داده‌های شیمی بکار می‌برد، یکی از اولین زمینه‌های تحقیق بوده است که در جهت تحلیل این گونه داده‌ها گامهایی برداشته است. بعضی از داده‌ها در این گرایش از شیمی، مشخصه‌های مهمی دارند که آنها را از سایر داده‌ها متمایز می‌کند. هر مشاهده به وسیله تعداد زیادی از متغیرهای اندازه‌گیری شده مشخص می‌شود که با یکدیگر همبستگی زیادی دارند. علاوه بر آن، ممکن است که تعداد این متغیرها از تعداد مشاهدات بیشتر باشد

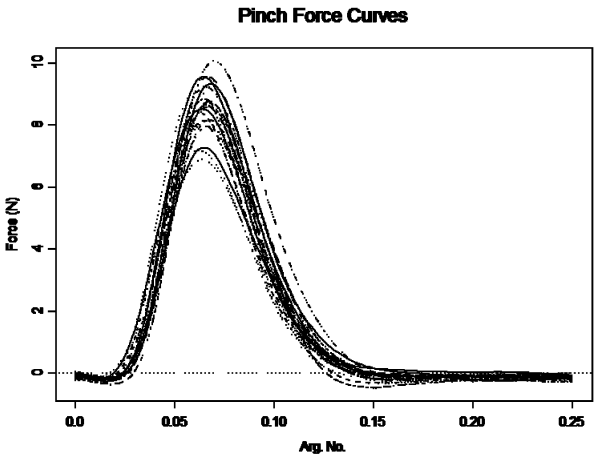
[۱] عضو هیأت علمی گروه آمار– دانشگاه شهید بهشتی– تهران– ایران

([۱۴]). برای مثال، اگر از اطلاعات طیفی انعکاس‌های اشعه‌ی ماورابنفش (NIR) برای بدست آوردن درصد چربی یا دیگر مواد سازنده در خمیر بیسکویت استفاده شود این اطلاعات برای یک سیگنال شامل صدها عدد و رقم می‌باشد. توابع تصادفی ممکن است یا به خودی خود مورد توجه باشند یا به خاطر اطلاعاتی که در مورد دیگر اندازه‌گیری‌ها به ما منتقل می‌کنند مورد علاقه واقع شوند. در این مطالعه، هدف می‌تواند کشف علل و چگونگی تغییرات در دما و استخراج الگوهای موجود در آن باشد. دسته‌ی دوم، دربرگیرنده‌ی مدل‌های خطی تابعی است که در آن داده‌های تابعی نقش متغیرهای توضیحی را بازی می‌کنند. برای مثال، با تشکیل یک مدل خطی تابعی می‌توان درصد چربی در بیسکویت را پیش‌بینی کرد. در این مثال، اطلاعات مربوطه به شکل نمایش ناپیوسته‌ای از یک سیگنال در طول موجهای مختلف که به عنوان شناسه این تابع تصادفی است مشاهده می‌شود. بنابراین در این مثال، متغیر پاسخ درصد چربی و متغیر مستقل $X(t)$، نشان دهنده‌ی شدت سیگنال در طول موج t است و هدف برآورد رابطه‌ی بین متغیر پاسخ (اسکالر) و متغیر مستقل (تابع تصادفی $X(t)$) می‌باشد ([۱۴]).

یکی از ویژگی‌های داده‌های تابعی آنست که تغییرات ناشی از اثر زمان را جدای از اثرات بوجود آمده از تفاوت‌های بین آزمودنی‌هایی (افراد یا اشیایی) که به تصادف از جامعه برای مطالعه انتخاب شده‌اند نشان می‌دهد. این خصوصیت، از طبیعت داده‌های جمع آوری شده ناشی می‌شود. داده‌های اولیه شامل اندازه‌گیری‌های مکرر از افراد، روی زمان است. بر خلاف مطالعات مقطعی [۲] که در آن یک کمیت برای هر فرد (شیء) اندازه‌گیری می‌شود در اینجا تحلیلگران قادر هستند تا هنگام تحلیل داده‌ها از ظرفیت آنها برای کشف اثر زمان استفاده نمایند. جدا سازی تغییراتی که در طول زمان و برای هر آزمودنی اتفاق می‌افتد از تغییرات میان آنها، برای آشکارسازی مشخصه‌های مفید جامعه بسیار سودمند می‌باشد. در یک نمونه مانند $X_1(t),...,X_n(t)$، تغییرات از نوع اول به متغیر t (زمان) مرتبط می‌شود که روی یک فاصله زمانی مانند $[a,b]$ اتفاق می‌افتد و تغییرات نوع دوم از طریق طبیعت تصادفی X قابل مشاهده است. مشاهده‌ی اثر متفاوت این دو نوع تغییر در قالب مثال زیر امکان‌پذیر می‌باشد.

یکی از مهم‌ترین گروههای ماهیچه‌ای در بدن انسان، انگشست شست و اشاره را در برداشتن یک شیء کنترل می‌کند. نیروی اعمال شده در برداشتن شیء باید با وزن و ساختار شیء متناسب باشد. در شکل (۱)، نیروی فشار در مقابل زمان رسم شده است. ۲۰ منحنی موجود در شکل (۱) بیانگر ۲۰ نوع فشار متفاوت می‌باشد. این مشاهدات از یک آزمایش بیومکانیک که به منظور تحقیق در مورد حرکت فشار دست طرح‌ریزی شده است بدست آمده‌اند. اگر t زمان مورد نظر باشد آنگاه $X(t)$ مقدار نیرو در زمان t را نشان می‌دهد. برای هر منحنی موجود در شکل (۱)، تغییرات در طول زمان نشان‌دهنده‌ی مقادیر متفاوت نیروی فشار در زمانهای مختلف است. امّا برای یک زمان خاص مانند $t = t_0$، تغییرات موجود در بین ۲۰ منحنی موجود نشان‌دهنده‌ی تغییرات تصادفی می‌باشد ([۱۹]).

[۲] cross-sectional



شکل ۱. بیست نیروی فشار متفاوت که در طول زمان ثبت شده‌اند([۱۹]).

برای کشف اثر زمان، می‌توان از ابزارهایی که برای بررسی رفتار توابع مورد استفاده قرار می‌گیرند سود برد. برای مثال، با استفاده از مشاهدات، قادریم تا مشتق مرتبه دوم هر تابع را برآورد کنیم. سپس با استفاده از این برآوردها، می‌توان مقدار انحنا یا نرخ رشد هر تابع را بدست آورد. بنابراین در مواردی که مشاهدات ذاتاً توابع پیوسته‌ای از زمان هستند تصور رکوردها (اندازه‌گیری‌های مکرر روی زمان) به عنوان یک تابع به جای در نظر گرفتن آنان به صورت یک بردار از مشاهدات در نقاط مجزای زمانی، ما را قادر می‌سازد تا حداقل با برآورد مشتقات توابع، از آنها برای بررسی اثر زمان در داده‌ها استفاده نماییم ([۱۸]).

تحلیل داده‌های تابعی را می‌توان با تحلیل داده‌های چندمتغیره قیاس کرد که در آن تعداد متغیرها (p) در حال افزایش به بی‌نهایت است ([۱۶]). بنابراین بردارها و ماتریس‌ها در فضاهای متناهی به توابع و عملگرها در فضاهای با بُعد بی‌نهایت تبدیل می‌شوند.

با توجه به بُعد بی‌نهایت در داده‌های تابعی، این داده‌ها ممکن است روشهای آماری متداول برای تحلیل داده‌ها را به چالش بکشند و در نتیجه برای تحلیل آنها انطباقهایی در نظریه‌های مربوطه لازم به نظر می‌رسد.

در این مقاله، ابتدا تعاریف و مفاهیم مربوط به مدل‌های خطی تابعی و برآورد ضرایب مدل و چگونگی بدست آوردن آن را بیان می‌کنیم. در ادامه با تحلیل یک مجموعه داده‌ی واقعی (داده‌های بارندگی و دما)، برآورد ضریب تابعی مجهول و مقادیر پیش‌بینی شده بدست آورده شده‌اند.

۲ مدل‌های خطی تابعی

فرض کنید که برای $j = ۱,\dots,p$ و $i = ۱,\dots n$، $X_i(t_j)$ نشان دهنده مقدار تابع پیوسته $X_i(.)$ در زمان t_j باشد که در آن $t_۱,...,t_p \in I$ نقاطی هستند که در آنها تابع پیوسته‌ی $X_i(t)$ مشاهده شده است. اگر یک مدل رگرسیون خطی با این متغیرهای مستقل مورد توجه قرار دهیم، برای تابعی مانند $b(t)$ می‌توانیم تابع خطی $E(Y_i|X_i) = \int_I X_i(t)\,b(t)\,dt$ را با سری $\sum_{j=۱}^{p} X_i(t_j)\,b(t_j)$ تقریب بزنیم. در این گونه مسایل، فرانک و فریدمن ([۶]) دو روش کمترین مربعات جزیی (PLS) و رگرسیون مؤلفه‌های اصلی (PCR) که عموما به وسیله شیمی سنج‌ها برای تحلیل این گونه داده‌ها مورد استفاده قرار می‌گیرند رابه طور خلاصه مورد بحث قراردادند. پس از آن مؤلفان، این دو روش را با رگرسیون ریدج (RR) در شرایط یکسان که در آن بردار ضرایب، مقید به قرار گرفتن در یک زیرفضا باشد و متغیرهای مستقل تصویر شده در این زیر فضا دارای واریانس بزرگتر باشند مورد مقایسه قرار دادند.

بحث هستی و ملوس ([۱۱]) در مورد این سه روش (PLS, RR, PCR) و همچنین روش کمترین مربعات تاوانیده[۳] کمک زیادی به فهم بهتر مسایلی که از تحلیل داده‌های تابعی نشات می‌گیرد ایجاد کرده است. هستی و ملوس تاکید نمودند که در حضور تعداد زیادی از متغیرهای کمکی $X_i(t_j)$، که معمولا تعدادشان بیشتر از حجم نمونه است، ممکن است مدل کارایی‌اش را برای پیش‌بینی از دست بدهد. علاوه برآن، همبستگی زیاد متغیرهای توضیحی $X_i(t_j)$ در مدل، منتهی به مدلی غیر قابل تفسیر می گردد. در ادامه، مؤلفان ذکر کردند چگونه می‌توان برآورد بهتری از شیب خط رگرسیون با مقید کردن آن از طریق روش کمترین مربعات تاوانیده بدست آورد. استفاده از این روش منجر به یک برآورد هموار برای ضریب رگرسیون می‌گردد. هستی و ملوس ([۱۱]) همچنین تاکید کردند که بهتر است از روشهایی که یک رابطه‌ی ترتیبی بین متغیرهای توضیحی در قالب مقادیر اندیس‌شان ایجاد می‌کنند استفاده شود. در این راستا، مسأله با استفاده از تعداد کمی از متغیرهای مستقل در مقایسه با صدها متغیر اولیه قابل حل است.

هستی و ملوس ([۱۱]) همچنین روش «بسط‌های پایه‌ی هموار» را برای مدل کردن شیب رگرسیون $b(t)$ به صورت هموار پیشنهاد دادند. در این روش، ضریب رگرسیون $b(t)$ در قالب بسطی از یک دنباله از توابع پایه‌ی هموار از قبیل چند جمله‌ای‌ها، توابع سینوسی و کسینوسی، اسپلاین و غیره نوشته می‌شود. در جواب به هستی و ملوس، فرانک و فریدمن ([٦]) تایید کردند که روشهایی که به رابطه‌ی ترتیبی در میان اندیس متغیرهای کمکی توجه می‌کنند و یا قیودی را برای هموواری ضریب $b(t)$ اعمال می‌کنند ممکن است بهتر از RR، PLS و PCR در حالتی که منحنی‌های $X_i(t)$ هموار نیستند عمل کنند. در حقیقت، چنین روش‌هایی که طبیعت تابعی مساله را به حساب می‌آورند در مقایسه با آنهایی که این کار را نمی‌کنند منطقی‌تر به نظر می‌رسند. در این زمینه می‌توان به ([۱۴]) که مزایای روش‌های تابعی را با روش‌های غیر تابعی از قبیل PLS و PCR مقایسه کرده‌اند مراجعه نمود.

کارهونن ([۱۳]) یک نظریه درباره‌ی فرآیندهای تصادفی در فضاهای هیلبرت ارائه داد. با استفاده از این نظریه، گرناندر ([۷]) توانست اولین گامها در جهت تحلیل داده‌های تابعی با استفاده از بسطهای کارهونن-لوی برای داده‌های تابعی بردارد. تلاش‌های او شامل یک طرح پیشنهادی برای رگرسیون تابعی نیز بود. امّا، مدلهای رگرسیون تابعی تنها بعد از کارهای اخیر [۱۷] و فصلهای ۱۵ و ١٦ از [۱۸] به طور گسترده مورد استفاده قرار گرفتند.

بسته به طبیعت متغیرهای پاسخ و متغیرهای مستقل، چهار نوع متفاوت از مدل‌های رگرسیون تابعی وجود دارد:

- متغیرهای پاسخ و مستقل به صورت تابع هستند.
- متغیر پاسخ یک تابع و متغیرهای مستقل به شکل بردار هستند.
- متغیر پاسخ اسکالر و متغیرهای مستقل تابع هستند.
- آنالیز واریانس تابعی ساده وقتی که متغیر پاسخ تابعی است.

[۳] Penalized Least Squares Method

مدلهایی از نوع اول به وسیله رمسی و دالزل ([۱۷]) معرفی شدند. این مدل‌ها را می‌توان به صورت تعمیم مدل رگرسیون خطی چند متغیره $E[Y|X] = \mathbf{B}\,\mathbf{X}$ که در آن $\mathbf{B}$ و $\mathbf{X}$ به ترتیب ماتریس ضرایب و ماتریس متغیرهای مستقل هستند در نظر گرفت. وقتی که متغیرهای پاسخ و مستقل به صورت توابع پیوسته ظاهر شوند، مدل به صورت زیر تبدیل می‌شود:

$$E[Y(t)|X] = \mu(t) + \int_I X(s)\,\beta(s,t)\,ds\,,$$

که در آن μ تابع میانگین پاسخ و β شیب رگرسیون است. سپس می‌توان پارامتر مجهول β را با استفاده از روشهایی از قبیل اسپلاین تاوانیده ([۱۲])، استفاده از بسط آن بر اساس یک پایه (نمایش پایه‌ای) یا بسطهای بریده (فصل ۱۶،[۱۸] و [۴]) برآورد کرد. همچنین در مورد مدلهایی که در آن پاسخ، یک تابع تصادفی و متغیرهای مستقل اسکالر یا بردار هستند در ([۴]) بحث شده است. نمونه‌هایی از مدلهایی که در آن متغیرهای مستقل تابع و پاسخ‌ها به صورت متغیرهایی از قبیل متغیرهای دودویی، نوع شمارشی یا نوع پیوسته هستند در کارهای ([۱], [۲] ، ([۳]، [۱۲] و [۱۵]) یافت می‌شود.

نوع سوم از مدلهای خطی، در مسایل پیش بینی تابعی به وجود می آید. برای مثال، می‌توان با بکارگیری این نوع مدل خطی، مقدار کل بارندگی در ایستگاه‌های هواشناسی (اسکالر) را با استفاده از الگوها و تغییرات دما در سراسر سال (تابع تصادفی) پیش‌بینی کرد. از دیگر موارد کاربرد این مدل رگرسیونی، می‌توان به استفاده از طیف‌بینی ماورابنفش برای بدست آوردن اطلاعات در مورد تعیین سطح یک پروتئین خاص در گونه های مختلف گندم اشاره نمود. فرض کنید $X_i(t)$، شدت ثبت شده از اشعه‌ی منعکس شده وقتی که طول موج برابر t و Y_i سطح یک پروتئین خاص برای i امین نوع گندم باشد. با تشکیل یک مدل رگرسیون خطی می‌توانیم مدل را با مشاهده جدید از$X(t)$ برای پیش‌بینی سطح پروتئین مورد نظر در گونه‌های گندم به کار برد ([۸]).

در مثال بالا، به دلیل مشکلاتی که در اندازه‌گیری متغیر وابسته وجود دارد پیش‌بینی مقدار متغیر وابسته با استفاده از مدل ساخته شده می‌تواند بسیار مفید باشد. اندازه‌گیری متغیر وابسته ممکن است با هزینه زیاد و با صرف زمان طولانی در یک آزمایشگاه انجام گیرد اما می‌توان متغیر مستقل $X_i(t)$ (اندازه های نورسنجی) را سریعتر و با هزینه کمتر با کمک یک ابزار خاص در سر مزرعه اندازه گرفت ([۸] و [۱۴]).

نوع چهارم از مدلهای خطی، در قالب آنالیز واریانس تابعی که در آن متغیر پاسخ و متغیر کمکی تابعی می‌باشند قابل بررسی است. مدل مربوطه به صورت زیر می‌باشد:

$$y_{sg}(t) = \eta(t) + \tau_g(t) + \varepsilon_{sg}(t), \quad \begin{matrix} s = ۱, ۲, .., N_g \\ g = ۱, ..., k \end{matrix} \quad (۱)$$

در این مدل تحت هر گروه N_g مشاهده قرار دارد و $y_{sg}(t)$ مقدار متغیر پاسخ برای sاُمین مشاهده تحت gاُمین گروه است. $\eta(.)$ میانگین کل می‌باشد و یک اثر کلی را برای تمام گروهها نشان می‌دهد. عبارت $\tau_g(.)$ اثر خاص گروه gاُم را روی متغیر پاسخ نشان می‌دهد. برای اینکه اثرات مذکور به طور منحصر به فرد برآورد شوند قید زیر را در نظر می‌گیریم:

$$\sum_g \tau_g(t) = ۰. \quad \forall t$$

فرض کنید ماتریس طرح $\boldsymbol{X}$ یک ماتریس $N \times (k+1)$ باشد که ستون gاُم آن تحت گروه $g+1$اُم بدست آمده و تمام درایه‌های ستون اول یک است. همچنین $N = \sum_g N_g$ و در هر سطر، درایه $g+1$اُم آن یک و بقیه صفر می‌باشند. بنابراین مدل (۱) را می‌توان به صورت زیر نوشت:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \varepsilon, \qquad (۲)$$

که $\boldsymbol{b}(.)$ برداری از ضرایب رگرسیونی $b_j(.)$ به صورت $b_1 = \eta, b_2 = \tau_1, ..., b_{k+1} = \tau_k$ است. اینک هدف برآورد پارامتر $\boldsymbol{b}$ به گونه‌ای است که معیار برآزش کمترین توانهای دوم مینیمم شود:

$$LMSSE(\boldsymbol{b}) = \int [\boldsymbol{y}(t) - \boldsymbol{X}\boldsymbol{b}(t)]'[\boldsymbol{y}(t) - \boldsymbol{X}\boldsymbol{b}(t)]dt.$$

با اضافه کردن عبارت تاوان هموواری به رابطه فوق، می‌توان برآوردها را بهبود بخشید؛ بنابراین برآوردها را برای این حالت بدست می‌آوریم. در این روش، هموواری روی پارامترهای $\boldsymbol{b}$ اعمال می‌شود (برای جزئیات بیشتر به ([۱۸]) مراجعه شود).

۱.۲ برآورد ضرایب در یک مدل خطی تابعی ساده

یک مدل خطی تابعی ساده به صورت:

$$Y_i = a + \int_I b(t)\, X_i(t)\, dt + \epsilon_i\,; \quad i = 1, ..., n, \qquad (۳)$$

است که در آن $b(.)$ و $X_i(.)$ توابعی از I به خط حقیقی و توان دوم آنها انتگرال‌پذیر می‌باشد، a، Y_i و ϵ_i اسکالر هستند، a و $b(.)$ غیر تصادفی و زوجهای (X_1, ϵ_1)و...و(X_n, ϵ_n) مستقل و همتوزیع‌اند. علاوه بر آن، توابع تصادفی X_i مستقل از خطاهای ϵ_i هستند، $\Sigma^2 = E(\epsilon^2) < \infty$، $E(\epsilon) = 0$ و $\int_I E(X(t)^2)\, dt < \infty$، که در آن ϵ و X همتوزیع با ϵ_i و X_i ها می‌باشند.

برآورد تابع $b(.)$ مسأله‌ای با بُعد بی‌نهایت است. در نتیجه در مدل خطی تابعی، حل این مسأله مستلزم استفاده از روشهای هموار کردن یا روشهای نظم می‌باشد که ما را قادر می‌سازد تا بُعد را کاهش دهیم. این مورد، وجه تمایز تحلیل مدل‌های خطی تابعی در مقایسه با تحلیل مدل‌های خطی کلاسیک است.

فرض کنید $X_1(t)$، $X_2(t)$،...، $X_n(t)$ یک نمونه به حجم n باشد آنگاه $\widehat{K}(u,v)$، به صورت زیر تعریف می‌شود:

$$\widehat{K}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\{X_i(u) - \bar{X}(u)\}\{X_i(v) - \bar{X}(v)\}\,,$$

که در آن $\bar{X}(.) = \frac{1}{n}\sum_{i=1}^{n} X_i(.)$. عملگر $\widehat{K}$ که به صورت

$$\left(\widehat{K}\widehat{\psi}\right)(x) = \int \widehat{K}(x,y)\,\widehat{\psi}(y)\,dy\,,$$

تعریف می‌شود، یک عملگر روی $L_2(I)$ است. همچنین داریم $(\widehat{K}\hat{\psi}_j)(t) = \hat{\theta}_j\hat{\psi}_j(t)$ که در آن $\hat{\theta}_j$ مقدار ویژه jاُم و $\hat{\psi}_j(t)$ تابع ویژه متناظر با آن است. اگر $X_i(.)$ و $b(.)$ را بر مبنای پایه متعامد $\widehat{\psi}_1(.)$ و $\widehat{\psi}_2(.)$ و... بنویسیم، آنگاه داریم:

$$X_i(t) = \sum_{j=1}^{\infty}\xi_{ij}\,\widehat{\psi}_j(t)\,, \qquad b(t) = \sum_{j=1}^{\infty}\bar{b}_j\,\widehat{\psi}_j(t)\,, \qquad (۴)$$

که در آن $\xi_{ij} = \int_I X_i(t)\,\widehat{\psi}_j(t)\,dt$ و $\bar{b}_j$ $\int_I b(t)\,\widehat{\psi}_j(t)\,dt$ به ترتیب نشان دهنده‌ی ضرایب

فوریه تعمیم یافته $X_i(.)$ و $b(.)$ می‌باشند. بنابراین، می‌توان مدل (۳) را به صورت زیر نوشت:

$$Y_i = a + \sum_{j=1}^{\infty} \bar{b}_j \, \xi_{ij} + \epsilon_i \, ; \qquad i = 1, ..., n. \qquad (۵)$$

هال و حسینی‌نسب ([۹] و [۱۰]) راهی برای غلبه بر مشکل بُعد بی‌نهایت با استفاده از مؤلفه‌های اصلی (PCA) ارائه دادند. در این روش، مؤلفان از تصویر کردن مشاهدات به داخل فضایی که به وسیله‌ی اولین r ویژه تابع $\widehat{K}$ به وجود می‌آید و منجر به یک نمایش بهینه برای $X_i(.)$ ها نسبت به واریانس توضیح داده شده می‌شود ([۵]) بهره بردند. بدین ترتیب، مؤلفه‌های غالب تغییرات توابع تصادفی، مدل رگرسیون خطی تابعی را به یک مدل رگرسیون معمولی با تعداد متناهی متغیر تصادفی مستقل (ξ_{ij}، $j = 1, \ldots, r$) کاهش می‌دهند. سپس پارامترهای مجهول، از طریق روش‌های متداول از قبیل روش‌های کمترین مربعات برآورد می‌شوند.

مقدار واقعی (a°, b°) از (a, b) به وسیله‌ی روشهای کمترین مربعات از طریق مینیمم کردن

$$S(a, \bar{b}_1, \ldots, \bar{b}_r) = \sum_{i=1}^{n} (Y_i - a - \sum_{j=1}^{r} \bar{b}_j \, \xi_{ij})^2 \,,$$

نسبت به a، $\bar{b}_1$،...، $\bar{b}_r$ و در نظر گرفتن $\bar{b}_j = 0$ برای $j \geq r+1$، برآورد می‌شود. می‌توان عبارت $S(a, \bar{b}_1, \ldots, \bar{b}_r)$ را به صورت $(\mathbf{Y} - \mathbf{\Lambda}^* \, \mathbf{b}^*)^T (\mathbf{Y} - \mathbf{\Lambda}^* \, \mathbf{b}^*)$ نوشت که در آن

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \; b^* = \begin{pmatrix} a \\ \bar{b}_1 \\ \vdots \\ \bar{b}_r \end{pmatrix},$$

$$\Lambda^* = \begin{pmatrix} 1 & \xi_{11} & \cdots & \xi_{1r} \\ \vdots & \vdots & & \vdots \\ 1 & \xi_{n1} & \cdots & \xi_{nr} \end{pmatrix}.$$

با مشتق گیری نسبت به $\mathbf{b}^*$ و برابر صفر قرار دادن آن، جواب ماکزیمم به صورت زیر است:

$$\begin{aligned} \hat{a} &= \bar{Y} - \sum_{j=1}^{r} \hat{b}_j \bar{\xi}_j \,, \\ \hat{b}_{(r)} &= (\hat{b}_1, \ldots, \hat{b}_r) = \widehat{\Sigma}_{(r)} \, \hat{g}_{(r)} \,, \qquad (۶) \end{aligned}$$

که در آن $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$، $\bar{\xi}_j = n^{-1} \sum_{i=1}^{n} \xi_{ij}$، $\widehat{\Sigma}_{(r)} = (\hat{\Sigma}_{jk})$ ماتریسی $r \times r$ و $\hat{g}_{(r)} = (\hat{g}_1, \ldots, \hat{g}_r)^T$. مؤلفه‌های ماتریس $\widehat{\Sigma}_{(r)}$ و بردار $\hat{g}_{(r)}$ به صورت زیر تعریف می‌شوند:

$$\begin{aligned} \hat{\Sigma}_{jk} &= \frac{1}{n} \sum_{i=1}^{n} (\xi_{ij} - \bar{\xi}_j)(\xi_{ik} - \bar{\xi}_k) \\ &= \int_I \widehat{K}(u, v) \, \widehat{\psi}_j(u) \, \widehat{\psi}_k(v) \, du \, dv = \hat{\theta}_j \, \delta_{jk} \,, \\ \hat{g}_j &= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(\xi_{ij} - \bar{\xi}_j) \,, \end{aligned}$$

که در آن δ_{jk} نشان دهنده‌ی دلتای کرونکر است. در نتیجه، $\widehat{\Sigma}_{(r)} = \mathrm{diag}(\hat{\theta}_1, \ldots, \hat{\theta}_r)$ و با استفاده از(۶) برای هر $u \in I$، برآوردگر b به صورت

$$\hat{b}(u) = \sum_{j=1}^{r} \hat{b}_j \, \widehat{\psi}_j(u) = \sum_{j=1}^{r} \hat{\theta}_j^{-1} \, \hat{g}_j \, \widehat{\psi}_j(u) \,,$$

است. می‌توان پارامتر هموار کننده‌ی r را از طریق اعتبارسنجی متقابل[۴] انتخاب نمود. این معیار برای داده های تابعی به صورت زیر معرفی شده است ([۹]):

$$CV(r) = \frac{1}{n} \sum_{j=1}^{n} \left(Y_i - \hat{a}_{-i;r} - \int_{\mathcal{I}} \hat{b}_{-i;r}(t) X_i(t) dt \right) ,$$

که در آن $(\hat{a}_{-i;r}, \hat{b}_{-i;r})$ برآورد کمترین توانهای دوم (a, b) است که بر اساس مجموعه داده های Z_i که شامل

[۴] Cross-validation

همه ی داده های (X_j, Y_j) بجز iامین زوج است محاسبه می شوند. در این معیار ابتدا با کنار گذاشتن iامین مشاهده، بر اساس $n-1$ مشاهده ی دیگر برآورد ضرایب انجام و سپس با استفاده از iامین زوج (X_i, Y_i) اعتبار برآوردهای بدست آمده ارزیابی می شود. این کار برای همه ی زوج مشاهدات، انجام و میانگین حاصل به عنوان یک معیار مورد توجه قرار می گیرد. انتخاب مقداری از r که به کمترین مقدار $CV(r)$ منجر شود یک انتخاب معقول برای r است.

۳ مقایسه مدل خطی چندگانه با مدل تابعی

مدل خطی تابعی شامل مشاهداتی از یک یا چند متغیر کمکی تابعی می‌باشد که متغیر پاسخ آن اسکالر یا تابعی است. در این بخش می‌خواهیم این نوع رگرسیون را با یک متغیر کمکی تابعی توضیح دهیم.
در حالتی که متغیر پاسخ اسکالر یا چندگانه است، مدل زیر را در نظر بگیرید:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \varepsilon.$$

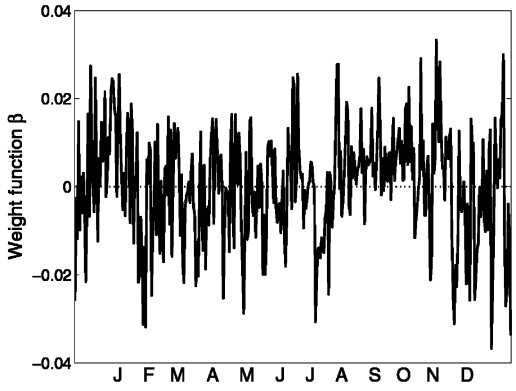
در این مدل، تبدیل خطی مورد نظر از فضای پارامتر به فضای مشاهدات بوسیله ماتریس طرح $\boldsymbol{X}$ تعیین می‌شود. یک شکل تابعی از مدل خطی این است که در آن متغیر پاسخ y_i اسکالر و متغیرهای توضیحی z_i تابعی باشند. به عنوان مثال می‌خواهیم مقدار کل بارندگی در سال برای ایستگاههای هواشناسی را بر اساس الگوی تغییرات دما پیش‌بینی کنیم. فرض کنید $\mathtt{Temp}_{ij}$ مقدار مشاهده شده دما در ایستگاه هواشناسی i اُم و در روز j اُم باشد. می‌خواهیم $\log \mathtt{Prec}_i$ (لگاریتم مقدار کل بارندگی سالانه برای ایستگاه i اُم) را بر اساس دما پیش‌بینی کنیم. لازم به ذکر است که به دلیل پراکندگی زیاد مقادیر بارندگی از لگاریتم آنها استفاده می‌کنیم. اگر بدون توجه به این که Temp تابعی است از روش رگرسیون معمولی برای تحلیل استفاده می‌کنیم. آنگاه مدل زیر را داریم:

$$\log Prec_i = \alpha + \sum_{j=1}^{365} Temp_{ij}\beta_j + \varepsilon_i \qquad i = 1, 2, ..., 35.$$

این مدل شامل ۳۵ معادله و ۳۶۶ مجهول است. حتی اگر ماتریس طرح پررتبه‌ی ستونی باشد دستگاه بی‌نهایت جواب دارد که هر کدام از جوابها یک مدل مناسب برای داده‌های مشاهده شده را نتیجه می‌دهد. شکل (۲) یکی از بی‌نهایت جواب برای ضرایب مدل را نمایش می‌دهد. امّا در عمل به دلیل یکتا نبودن جواب، استفاده از چنین روشی سودمند نیست. این اشکال به دلیل آن پدید می آید که با استفاده از گسسته‌سازی بی‌نهایت مقدار (پارامتر) از تابع $\beta(t)$ به وجود می آید امّا تعداد معادلات موجود

$$y_i = \alpha + \int Temp_i(t)\beta(t)dt \qquad (7)$$

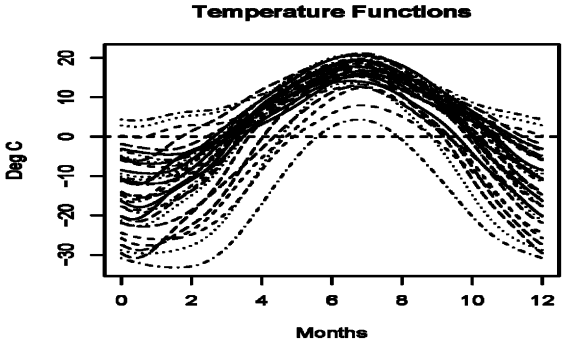
متناهی است که بر اساس آنها برآورد صورت می‌گیرد. پس بهتر است از مدل خطی تابعی استفاده کنیم ([۱۸]).



شکل ۲. تابع وزنی β برای پیش‌بینی کل لگاریتم بارندگی سالانه از الگوی مشاهدات سالانه دما.

۴ تحلیل مدل داده‌های دما و بارندگی

داده‌های واقعی شامل متوسط دمای روزانه بر حسب درجه سانتی‌گراد و مقدار کل بارندگی در طی سال است که از ۳۵ ایستگاه هواشناسی متفاوت در کانادا در سال ۱۹۸۲ جمع آوری شده است. نمودار ۳۵ تابع دما بر حسب زمان در شکل ۳ ارائه شده است. لازم به ذکر است که تحلیل این داده ها با استفاده از نرم افزار Splus انجام شده است. علاقمندیم تا تغییرات کل مقدار بارندگی سالانه را با استفاده از الگوهای تغییرات دما در سراسر کانادا بیان کنیم. با در نظر گرفتن مدل (۷)، فرض می‌کنیم y، مقدار بارندگی و $\mathrm{Temp}(t)$ دما باشد. به دلیل آن که، مقدار کل بارندگی مربوط به چهار منطقه‌ی متفاوت از سراسر کانادا می‌باشد مقدار آن در میان ایستگاه های هواشناسی با تغییرات زیادی همراه است. بنابراین، از لگاریتم آن به عنوان متغیر وابسته استفاده می‌کنیم. پارامتر هموار کننده‌ی r نیز با استفاده از معیار اعتبارسنجی متقابل $\hat{r} = ۲$ بدست آمده است و به کمک آن شیب رگرسیونی $\beta(t)$ و نقطه ثابت α را بر اساس فرمول (٦) برآورد کرده‌ایم.



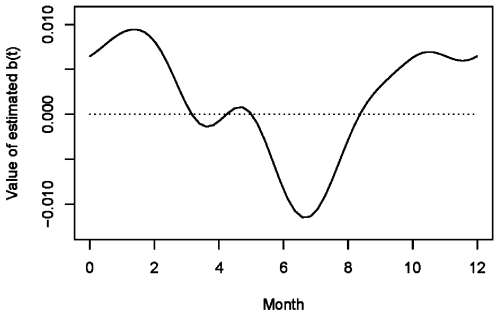
شکل ۳. دمای کانادا در سال ۱۹۸۲ که از ۳۵ ایستگاه هواشناسی مختلف در سراسر کانادا جمع آوری شده‌اند.

همچنان که شکل (۴) نشان می‌دهد شیب برآورد شده‌ی مدل خطی تابعی، وزنهای بیشتر را به ماههای زمستان و نیمه دوم پاییز اختصاص می‌دهد. علاوه بر آن، وزنهای کمتر متناظر با تابستان و نیمه بهار می‌باشند. همچنین بر اساس $\hat{r} = ۲$ و برآوردهای $\hat{\beta}(t)$ و $\hat{\alpha}$، می‌توان میزان بارندگی را بر اساس مدل برآورد شده

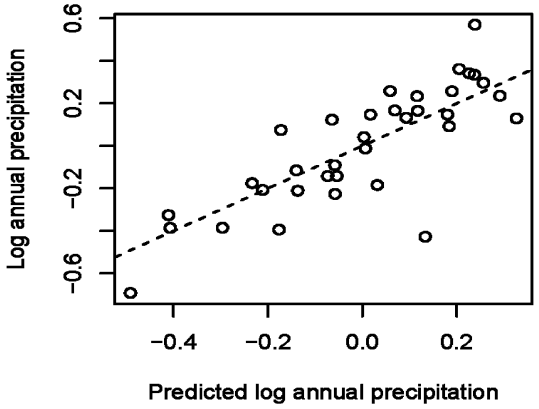
$$\widehat{\log Prec_i} = \hat{\alpha} \quad + \quad \int_I \hat{\beta}(t)\, Temp_i(t)\, dt\,; \qquad i = ۱, ..., n, \qquad (۸)$$

پیش‌بینی کرد. شکل (۵) نمودار لگاریتم مقادیر مشاهده شده را در برابر مقادیر پیش‌بینی شده لگاریتم بارندگی نشان می‌دهد. این نمودار نشان می‌دهد که مقادیر پیش‌بینی شده بر اساس مدل برآوردشده (۸) نسبتاً قابل قبول می‌باشند. لازم به ذکر است که برای سنجش نیکویی برازش مدل در حالت تابعی معیارهایی مشابه با

حالت رگرسیون خطی کلاسیک وجود دارد. برای مطالب بیشتر دراین مورد می توان به [۱۸] مراجعه نمود.



شکل ۴. شیب برآورد شده رگرسیون وقتی که متغیر وابسته لگاریتم کل مقدار بارندگی سالانه است. پارامتر هموارکننده $\hat{r} = ۲$ و نقطه‌ی ثابت $\hat{\alpha} = -۰.۲۴۱۳$ برآورد شده است.



شکل ۵. نمودار مقادیر مشاهده شده (لگاریتم مقدار بارندگی کانادا در سال ۱۹۸۲) در برابر مقادیر مقادیر پیش بینی شده. خط راست روی نمودار، نشان دهنده مقدار مشاهده شده=مقدار پیش بینی شده می‌باشد.

مراجع

[1] Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *statist. Probab. Lett.* **45**, 11-22.

[2] Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica*, **13**, 571-591.

[3] Cardot, H., Ferraty, F., Mas, A. and Sarda, P. (2003). Testing hypotheses in the functional linear model. *Scand. J. Statist.* **30**, 241-255.

[4] Chiou, J. M., Muller, H-G. and Wang, J. L. (2004). Functional Respone Models. *Statistical Sinica*. **14**, 675-693.

[5] Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some allications to statistical inference. *J. Multivariate Anal.* **12**, 136-154.

[6] Frank I. E. and Friedman J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*. **35**, 109-135.

[7] Grenander, U. (1950). Stochastic process and statistics inference. *Arkiv för Mstematik*, 196-276.

[8] Hall, P. and Horowitz, J. L. (2006). Methodology and convergence rates for functional linear regression. Manuscript.

[9] Hall, P. and Hosseini-nasab, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **68**, pp 109-126.

[10] Hall, P. and Hosseini-Nasab, M. (2009). Theory for high-order bounds in functional principal components analysis. Accepted by *Mathematical Proceedings of Cambridge Philosophical Society* .

[11] Hastie, T. and Mallows, C. (1993). A discussion of "A statistical view of some chemometrics regression tools" by I. E. Frank and J.H. Friedman. *Technometrics*. **35**, 140-143.

[12] James, G. M. (2002). Generalized linear models with functional predictors. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **64**, 411-432.

[13] Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae.* **A I 37**.

[14] Marx, B. D. and Eilers P. H. (1999). Generalized linear regression on sampled signals and curves: a P-spline approch. Technometrics. **41**, 1-13.

[15] Muller H.-G. and Stadmuller, U. (2005). Generalized functional linear models. *Ann. Statist.* **33**, 774-805.

[16] Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, **47**(4), 379-396.

[17] Ramsay J. O. and Dalzell, C. J. (1991). Some tols for functional data analysis (with discussion). *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **53**, 539-572.

[18] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis, 2nd Edition.* Springer, New York.

[19] Ramsay, J. O., Wang, X. and Flanagan, R. (1995). The Functional Data Analysis of Pinch Force. *Statist*, **44**(1), 17-30.