تحلیل داده‌های تابعی: تعمیمی از تحلیل داده‌های چندمتغیره

نوراللّه تازیکه‌میاندره [۱]، نصیبه خیراللّه‌زاده [۲]، هادی موقری [۳] سیدمحمدابراهیم حسینی‌نسب [۴]

چکیده:

بسیاری از پدیده‌ها مانند دما و میزان بارندگی، ذاتاً توابع پیوسته‌ای از زمان هستند. هر چند که با توجه به روش جمع‌آوری این گونه داده‌ها، استفاده از روش‌های چندمتغیره کلاسیک برای تحلیل آنها امکان‌پذیر است ولی این تحلیل با مشکلاتی مواجه است. ماهیت تابعی آن‌ها موجب می‌شود که این داده‌ها به جای مورد توجه قرار گرفتن در فضاهای برداری با بُعد متناهی، در فضاهای تابعی با بُعد نامتناهی در نظر گرفته شوند. بر این اساس انجام انطباق‌هایی در تئوری‌های مرتبط با مشاهدات چندمتغیره برای مناسبت استفاده از آن‌ها در تحلیل داده‌های تابعی لازم به نظر می‌رسد. چنین اقدامی به رویکرد جدیدی در آمار منتهی می‌شود که تحلیل داده‌های تابعی [5] نامیده می‌شود [٦] . در این مقاله تعاریف و مفاهیم مربوط به تحلیل داده‌های تابعی و ضرورت استفاده از آن، چگونگی تبدیل داده‌های گسسته‌ی اولیه به منحنی‌های پیوسته، مقایسه داده‌های تابعی با داده‌های چندمتغیره و طولی به همراه آماره‌های مکانی و پراکندگی تابعی ارائه خواهد شد. همچنین در ادامه به تحلیل مؤلفه‌های اصلی تابعی اشاره می‌کنیم.

واژه‌های کلیدی: تحلیل داده‌های تابعی، تحلیل داده‌های چندمتغیره، تحلیل مؤلفه‌های اصلی تابعی.

۱ مقدمه

در حالت کلی مشاهداتی که ذاتاً تابع پیوسته‌ای از یک متغیر مانند زمان یا فرکانس هستند، داده‌های تابعی نامیده می‌شوند. در عمل مشاهدات اولیه مربوط به هر آزمودنی اکثراً در تعدادی نقاط مجزا در طول یک بازه‌ی زمانی اندازه‌گیری و گزارش می‌شوند؛ بنابراین تحلیل چنین داده‌هایی را می‌توان با استفاده از روش‌های موجود در تحلیل داده‌های چندمتغیره انجام و تفسیر نمود. امّا با انجام تحلیل چندمتغیره برای داده‌هایی با ماهیتی ذاتاً تابعی ۱– امکان کنکاش بیشتر در مورد تغییرات در طول زمان (مثلاً برآورد و بررسی مشتق مراتب اول و دوم) از ما سلب می‌شود. ۲– وابستگی مشاهدات جمع‌آوری شده از هر آزمودنی در نقاط مجزا (وابستگی‌های افقی [٦]) نادیده گرفته می‌شود. در نتیجه جابه‌جایی ترتیب زمان‌های جمع‌آوری داده‌ها (متغیرها)، که هم ارز تغییر مرتبه‌ی اندیس‌ها در یک بردار چند مؤلفه‌ای می‌باشد، نتایج تحلیل آماری را تغییر نمی‌دهد. ۳– اطلاعات

---

[۱] گروه آمار– دانشگاه تربیت مدرس
[۲] گروه آمار– دانشگاه تربیت مدرس
[۳] گروه آمار– دانشگاه تربیت مدرس
[۴] گروه آمار– دانشگاه تربیت مدرس
[5] Functional Data Analysis(FDA)
[٦] Horizontal dependencies

مربوط به صفت مورد اندازه‌گیری در بین نقاط نادیده گرفته می‌شود. ۴- با افزایش تعداد متغیرها (p)، بُعد ماتریس واریانس-کوواریانس بدون در برداشتن اطلاعات مفیدی در مورد تغییرات بین مؤلفه‌ها با سرعت افزایش می‌یابد؛ زیرا تغییرات بین مؤلفه‌ها در این حالت بسیار کوچک است.

امّا در تحلیل داده‌های تابعی ، داده‌ها با در نظر گرفتن ماهیّت واقعی آن‌ها یعنی توابع پیوسته مورد تحلیل قرار می‌گیرند و اثر زمان در کل بازه‌ی زمانی و نه لزوماً در زمان‌های گسسته یا فواصل زمانی برابر، مورد توجه می‌باشد. بنابراین استفاده از تحلیل داده‌های تابعی در بررسی و تفسیر داده‌های تابعی تنها راه صحیح در پیش روی تحلیل‌گر می‌باشد [٦].

تحلیل داده‌های تابعی در چند سال اخیر با رشد فزاینده‌ای همراه بوده است و متون و مقالات زیادی مرتبط با این گرایش وجود دارد که از آن جمله می‌توان به [۴] و [٦] اشاره نمود.

در بخش (۲)، به چگونگی تبدیل داده‌های گسسته‌ی اولیه به منحنی‌های پیوسته پرداخته می‌شود. در ادامه نیز بعد از مقایسه داده‌های تابعی با داده‌های چندمتغیره و طولی در بخش‌های (۳) و (۴)، تعریف آماره‌های خلاصه مکانی و پراکنش تابعی ارائه می‌شود. آن‌گاه در ادامه به تحلیل مؤلفه‌های اصلی تابعی اشاره می‌کنیم.

۲ تبدیل داده‌های گسسته اولیه به منحنی‌های پیوسته

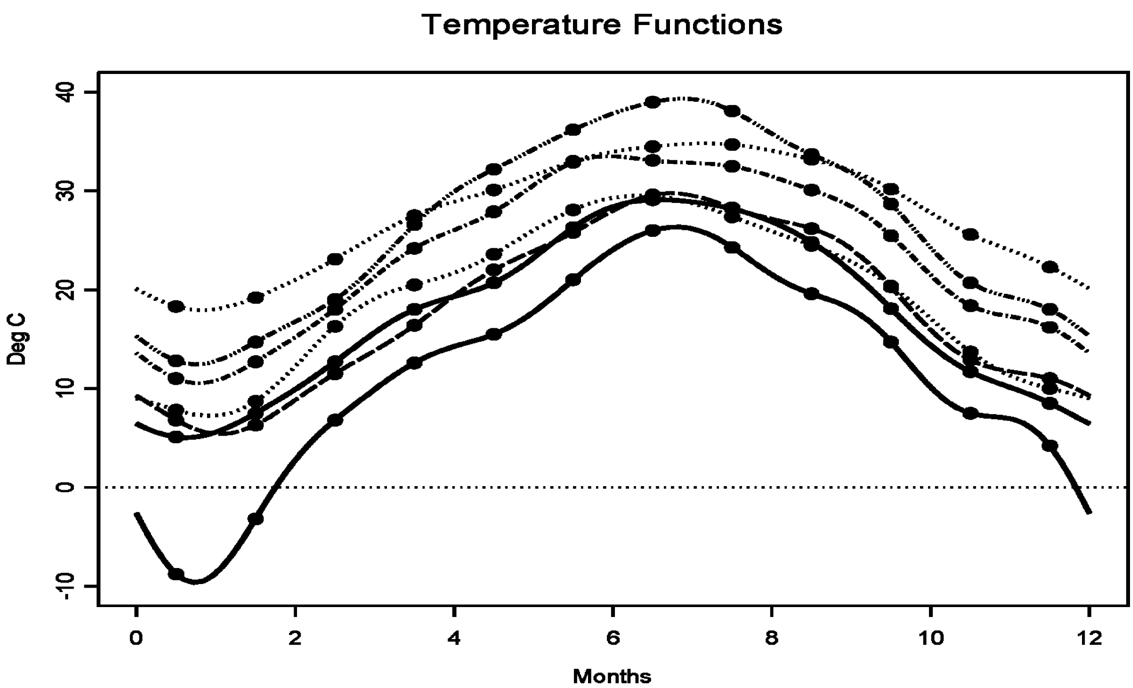
یکی از تفاوت‌های اصلی بین تحلیل داده‌های تابعی و تحلیل چندمتغیره ارزیابی هر مشاهده تابعی به عنوان یک ماهیت واحد و نه به صورت دنباله‌ای از مشاهدات گسسته می‌باشد. از طرف دیگر، به دلیل آن که در عمل مشاهدات اولیه به صورت گسسته در دسترس هستند؛ باید این مشاهدات گسسته را به مشاهدات تابعی تبدیل کنیم. با توجه به امکان وجود خطای اندازه‌گیری هنگام جمع آوری مشاهدات، بهتر است که این کار با استفاده از روش‌های هموارسازی صورت پذیرد. به بیانی دقیق‌تر، چنانچه مشاهدات بدون خطا اندازه‌گیری شده باشند کافیست از روش‌های درون یابی استفاده شود. امّا با توجه به وجود عوامل غیر قابل کنترل مانند خطاهای انسانی و ابزاری، معمولاً اعداد گزارش شده با خطا همراه هستند. از این رو، لازم است تا برای تحلیل آن ها از روش‌های هموارسازی استفاده شود.

۱.۲ استفاده از روش‌های هموارسازی در تحلیل داده‌های تابعی

برازش منحنی‌های مناسب به داده‌های گسسته با استفاده از روش‌های هموارسازی از طرق گوناگونی از جمله اسپلاین‌ها[۷]، اسپلاین‌های هموارسازی[۸]، هموارسازهای چندجمله‌ای موضعی[۹] انجام می‌شود. همچنین استفاده از توابع پایه برای برازش منحنی‌ها از دیگرروش‌های

[۷] Splines
[۸] Smoothing spline
[۹] Local polynomial smoothing

هموارسازی معمول می‌باشد. هر تابع را می‌توان به صورت مجموع وزنی یا ترکیب خطی از G تابع پایه نوشت. این روش منجر به سادگی محاسبات می‌شود و به این دلیل کاربردهای فراوانی دارد [٦]. به دلیل آن‌که پدیده‌ی دما دوره‌ای می‌باشد؛ در برازش منحنی به مشاهدات هر ایستگاه از بسط سری‌های فوریه استفاده شده است (برای بحث‌های تکمیلی به [۷] و [۱] مراجعه شود). در نهایت هر یک از توابع بدست آمده به این طریق، به عنوان یک مشاهده منفرد در تحلیل داده‌های تابعی مورد بررسی قرار خواهد گرفت [٦].



شکل ۱. دمای ایران در طول سال ۲۰۰۵ که از ۷ ایستگاه هواشناسی مختلف اندازه‌گیری شده است. روی هر منحنی ۱۲ نقطه نشان داده شده است که میانگین ماهیانه دما برای ایستگاه متناظر می‌باشند و به هر کدام از ۱۲ نقطه با استفاده از روش‌های هموارسازی یک منحنی هموار برازش شده است.

شکل (۱) دمای ایران در سال ۲۰۰۵، که از ۷ ایستگاه هواشناسی مختلف گزارش شده است را نمایش می دهد. داده‌های اولیه برای دما از هر ایستگاه، به صورت ۱۲ عدد که میانگین ماهیانه دما می‌باشند، گزارش شده است. با توجه به این که دما ذاتاً تابعی پیوسته از زمان است و می‌توان آن را برای هر نقطه از زمان مورد بررسی قرار داد؛ از این رو، مشاهدات مربوط به این ایستگاه‌ها را در قالب یک نمونه ۷ تایی از منحنی‌ها (توابع)، که هر کدام مربوط به یک ایستگاه است، در نظر می‌گیریم. به دلیل آن که داده‌های اولیه برای هر ایستگاه، ۱۲ مقدار گسسته را در بر می‌گیرد؛ نخستین گام در تحلیل آن ها به روش تابعی، برازش منحنی‌های مناسب به این داده‌های گسسته می‌باشد به طوری که هر منحنی از ۱۲ نقطه متناظر با آن عبور کند. بدین ترتیب مقدار دما در طول سال

مشخص است. از طرف دیگر، مشاهدات اولیه ممکن است با خطا اندازه‌گیری شوند، از این رو لازم می شود در هنگام برازش منحنی‌ها به نقاط، از روش‌های هموارسازی استفاده گردد تا خطا در حد امکان فیلتر شود.

۳ مقایسه داده‌های تابعی با داده‌های چندمتغیره

برای فهم بهتر ماهیت داده‌های تابعی و روش‌های تحلیل آن‌ها، می‌توان این داده‌ها را با داده‌های چندمتغیره مقایسه نمود. در واقع توجیه فضاهایی که توابع پیوسته‌ی حقیقی-مقدار به آن تعلق دارند با فضاهای برداری که مشاهدات بردار-مقدار به آن تعلق دارند امکان‌پذیر است. این موضوع سبب می‌شود، داده‌های تابعی را به عنوان تعمیمی از داده‌های چندمتغیره در نظر بگیریم.

در تحلیل داده‌های چندمتغیره، ماتریس مشاهدات به صورت

$$\boldsymbol{X} = \begin{pmatrix} x_{۱۱} & \cdots & x_{۱p} \\ \vdots & \ddots & \vdots \\ x_{n۱} & \cdots & x_{np} \end{pmatrix} \qquad (۱)$$

است که در آن مؤلفه‌ی x_{ij} نشان دهندهٔ مقدار متغیر j اُم برای آزمودنی i اُم می‌باشد. بنابراین مشاهدات به صورت بردارهای $\boldsymbol{x}_i = (x_{i۱}, \cdots, x_{ip})^t; \quad i = ۱, ۲, \cdots, n$ می‌باشند که در آن $\boldsymbol{x}_i \in I\!R^p$. امّا در تحلیل داده‌های تابعی ، مقدار متغیر پاسخ برای آزمودنی i اُم به صورت $x_i(t)$، $t \in \mathrm{I} = [۰, T]$، نمایش داده می‌شود. با وجود این تفاوت، با اقتباس از نگاشت‌های ماتریسی در حالت چندمتغیره، نگاشت‌های تابعی برای مشاهداتی با بُعد بی‌نهایت قابل تعریف هستند [۳].

در عمل، پدیده‌های تابعی معمولاً در زمان‌های گسسته، $۰ < t_۱ < t_۲ < \cdots < t_p < T$، اندازه‌گیری می‌شوند؛ از این رو، فرض کنید عناصر x_{ij} در ماتریس $\boldsymbol{X}$، مقدار مشاهده شده $x_i(t)$ در نقطه‌ی $t = t_j$ باشد. در این صورت، در عمل توابع تصادفی $x_۱(t), \cdots, x_n(t)$ در نقاط $t_۱, \cdots, t_p$ مشاهده شده‌اند. حال اگر $p \to \infty$ آنگاه ستون‌های ماتریس $\boldsymbol{X}$ در حال افزایش است. در این حالت مقدار هر $x_i(t)$ در نقاط مختلف t قابل مشاهده شدن است که این مفهوم گذر از داده‌های چندمتغیره (بردارها) و رسیدن به داده‌های تابعی (توابع) می‌باشد. با این استدلال، داده‌های تابعی در واقع تعمیمی از داده‌های چندمتغیره می‌باشند که در آن تعداد متغیره(p) در حال افزایش به بی‌نهایت است [۳].

تفاوت اصلی بین فضاهای برداری متناهی و نامتناهی در این است که عناصر یک فضای برداری متناهی را می‌توان براساس مجموع وزنی تعداد متناهی از بردارهای پایه بیان نمود ولی در فضاهای با بُعد نامتناهی چنین موضوعی لزوماً صادق نیست. اکثر فضاهای تابعی نیز در زمره‌ی فضاهای نامتناهی قرار می‌گیرند. از این رو، نمی‌توان عناصر آنها را به صورت ترکیب خطی متناهی از عناصر پایه نوشت.

ماتریس $\boldsymbol{X}$ در رابطهٔ (۱) را می‌توان به عنوان یک نگاشت از یک فضای برداری به فضای برداری دیگر مورد بررسی قرار داد [۳]. این دو فضای برداری عبارتند از:

۱) **فضای آزمودنی‌ها** (E) **با بُعد** p. هر آزمودنی در این فضا را می‌توان با یک نقطه نشان داد که

موقعیت این نقطه، متناظر با رابطه‌اش با هر یک از p متغیر می‌باشد. در این فضا، حاصلضربِ داخلی دو بردار $\boldsymbol{e}_j$ و $\boldsymbol{e}_k$ به صورت $b_E(\boldsymbol{e}_j,\boldsymbol{e}_k)=\boldsymbol{e}_j^t\boldsymbol{e}_k$ نمایش داده می‌شود که در آن $\boldsymbol{e}_j^t$ نشان دهندهٔ ترانهادهٔ بردار $\boldsymbol{e}_j$ می‌باشد[10]. این فضا را می‌توان با یک مجموعه‌ی p تایی از بردارهای یکامتعامد[11] تولید کرد.

۲) **فضای متغیرها** (F) **با بُعد** n. هر متغیر در این فضا را می‌توان با مقادیری که به هر یک از n آزمودنی در آن موقعیت اختصاص می‌دهد بیان کرد. در این فضا نیز یک ضرب داخلی $b_F(\cdot,\cdot)$ و یک مجموعه از n بردار مولد یکامتعامد وجود دارد.

بنابراین، به‌طور کلی ماتریس $\boldsymbol{X}_{(n\times p)}$ نگاشت‌های زیر را تعریف می‌کند:

۱) $\boldsymbol{X}:E\longrightarrow F$. فرض کنید $\boldsymbol{e}$ یک عضو از E باشد؛ آنگاه $\boldsymbol{f}_{(n\times 1)}=\boldsymbol{X}\boldsymbol{e}$، که موقعیتی را در فضای متغیرها نشان می‌دهد.

۲) $\boldsymbol{X}^t:F\longrightarrow E$. برای هر عنصر $\boldsymbol{f}\in F$، حاصلضرب $\boldsymbol{X}^t\boldsymbol{f}$ عضوی از E می‌باشد. این نگاشت، نگاشت ترانهاده[12] متناظر با ماتریس $\boldsymbol{X}$ نامیده می‌شود.

به‌طور کلی، هر نگاشت و نگاشت ترانهاده‌اش در رابطه زیر صدق می‌کنند:

$$b_F(\boldsymbol{f},\boldsymbol{X}\boldsymbol{e})\;=\;b_E(\boldsymbol{X}^t\boldsymbol{f},\boldsymbol{e}).$$

علاوه بر دو نگاشت فوق‌الذکر متناظر با ماتریس $\boldsymbol{X}$، دو نگاشت مهم دیگر را نیز می‌توان مورد بررسی قرار داد. به دلیل آن‌که این فضاها، نگاشت‌هایی از یک فضا در خودشان معرفی می‌کنند عملگر[13] نامیده می‌شوند:

۳) $\boldsymbol{V}:E\longrightarrow E$. با توجه به اینکه $\boldsymbol{X}$ و $\boldsymbol{X}^t$، به ترتیب یک بردار را از E به F و F به E نگاشت می‌کنند؛ از این رو، تلفیق $\boldsymbol{X}^t\circ\boldsymbol{X}$، فضای E را در خودش نگاشت می‌کند. به بیان ساده‌تر، اگر قرار دهیم $\boldsymbol{V}=\boldsymbol{X}^t\circ\boldsymbol{X}$؛ این نگاشت متناظر با ضرب $\boldsymbol{V}\boldsymbol{e}$ می‌باشد که در آن $\boldsymbol{e}\in E$.
لازم به ذکر است وقتی که میانگین هر ستون در ماتریس $\boldsymbol{X}$ صفر است؛ ماتریس $n^{-1}\boldsymbol{V}$، همان ماتریس واریانس-کوواریانس نمونه می‌باشد. در ضمن، اگر E و F هر دو فضای تابعی باشند آنگاه $n^{-1}\boldsymbol{V}$ را عملگر واریانس-کوواریانس نامند.

۴) $\boldsymbol{W}:F\longrightarrow F$. مشابه با عملگر $\boldsymbol{V}$ برای فضای E، عملگری نیز برای F وجود دارد که با اعمال $\boldsymbol{X}^t$ روی عنصر $\boldsymbol{f}\in F$ و سپس اثر دادن $\boldsymbol{X}$ روی تصویر $\boldsymbol{f}$ در E حاصل می‌شود. به عبارت دیگر این عملگر به صورت $\boldsymbol{W}=\boldsymbol{X}\circ\boldsymbol{X}^t$ است و متناظر با ضرب $\boldsymbol{W}\boldsymbol{e}$ می‌باشد.

[10] اگر روی یک فضای برداری یک ضرب داخلی (Inner product) تعریف شود آنگاه آن فضا را، فضای ضرب داخلی (Inner product space) گویند. ضرب داخلی، یک تابع حقیقی مقدار با سه ویژگی تقارن (Symmetry)، مثبت‌بودن (Positivity) و خطی در آرگومان (Bilinearity) می‌باشد. بطور مثال در یک فضای برداری که عناصر آن توابع پیوسته باشند؛ انتگرالِ ضربِ دو تابع، یک ضرب داخلی را با سه ویژگی فوق‌الذکر، تعریف می‌کند.
[11] Orthonormal vectors
[12] Transpose mapping
[13] Operator

می‌توان فضاهای E و F را به عنوان فضاهایی از توابع در نظر گرفت. برای مثال، F را می‌توان به عنوان فضای نگاشت‌های ممکن از n آزمودنی به اعداد حقیقی در نظر گرفت. آنگاه می‌توان مسأله را در قالب تابعی و در فضای توابع بیان نمود.

بحث‌های فوق برای موقعیتی که در آن داده‌ها برای آزمودنی iاُم، به شکل یک تابع یعنی $x_i(t), \quad 0 \le t \le T$ می‌باشند به صورت زیر است. در این حالت $\boldsymbol{X}$، دیگر یک ماتریس نیست بلکه مفهوم یک نگاشت را دارد. در واقع $\boldsymbol{X}$ نگاشت‌هایی را در فضاهای زیر تعریف می‌کند:

۱) **فضای آزمودنی‌ها با بُعد بی‌نهایت.** هر آزمودنی را می‌توان با یک تابع $e(t)$ متناظر کرد. چون هر تابع را می‌توان به صورت یک p بردار که در آن p در حال افزایش است تصور کرد، این فضا دارای بُعد بی‌نهایت می‌باشد. ضرب داخلی $b_E(\boldsymbol{e}_j, \boldsymbol{e}_k)$ برای توابع $e_j(t)$ و $e_k(t)$ به صورت زیر تعریف می‌شود:

$$b_E(\boldsymbol{e}_j, \boldsymbol{e}_k) = \int_0^T e_j(t) e_k(t) dt.$$

لازم به ذکر است چنانچه خود را به فضای هیلبرت با پایه‌ی شمارا مثل $L_2([0, T])$ محدود کنیم؛ هر عنصر را می‌توان بر حسب مجموع وزنی از تعداد شمارا از توابع یکامتعامد بیان کرد.

۲) **فضای زمانی F با بُعد** n. این فضا تمام مشخصات مربوط به فضای F در حالت pـمتغیره را در خود دارد. هر نقطه در این فضا با مقادیر n تابع در آن نقطه‌ی زمانی قابل بیان است.

در نهایت مشابه با حالت چندمتغیره، نگاشت‌هایی که $\boldsymbol{X}$ معرفی می‌کند عبارتند از:

۱) $\boldsymbol{X} : E \longrightarrow F$. فرض کنید $e(t)$ یک عضو از E باشد. آنگاه بردار $\boldsymbol{f}$ که iاُمین عنصرش به صورت زیر محاسبه می‌شود؛ عنصری در فضای n بُعدی و بنابراین عضوی از F است:

$$f_i = \int_0^T x_i(t) e(t) dt. \qquad (2)$$

لازم به ذکر است که این مسأله با آنچه که در حالت pـمتغیره بدست آمده، متناظر می‌باشد که در آن مجموع برای عنصر iاُم در حاصلضرب $\boldsymbol{Xe}$ با انتگرال تعویض شده است.

۲) $\boldsymbol{X}^t : F \longrightarrow E$. فرض کنید $\boldsymbol{f}$ یک بردار از F باشد؛ آنگاه تابع

$$\boldsymbol{X}^t : \boldsymbol{f} \longrightarrow \boldsymbol{X}^t \boldsymbol{f} = e(t) = \sum_{i=1}^{n} f_i \times x_i(t), (3)$$

عنصری در فضای E خواهد بود. به علاوه،به سادگی می‌توان نشان داد که برای هر $e(t) \in E$ و $\boldsymbol{f} \in F$ رابطه‌ی $b_F(\boldsymbol{Xe}, \boldsymbol{f}) = b_E(\boldsymbol{e}, \boldsymbol{X}^t \boldsymbol{f})$ برقرار است.

۳) $\boldsymbol{V} : E \longrightarrow E$. در این حالت نیز، یک تابع مانند $\boldsymbol{e}(t)$ از E با استفاده از (۲) بتوی F و آنگاه نتیجه‌ی نگاشت بردار nـبُعدی حاصل با استفاده از (۳) به E متعلق می‌باشد. این نگاشت را با عملگر $V = \boldsymbol{X}^t \circ \boldsymbol{X}$ می‌توان نشان داد که همانند یک ماتریس متقارن با مرتبه‌ی بسیار بزرگ (بی‌نهایت) می‌باشد:

$$\boldsymbol{X}^t \circ \boldsymbol{X} : \boldsymbol{e} \longrightarrow \boldsymbol{Ve}$$

$$= \sum_{i=1}^{n} x_i(t) \left[\int_{0}^{T} x_i(u) e(u)\, du \right],$$

$$= \int_{0}^{T} \left[\sum_{i=1}^{n} x_i(t) x_i(u) \right] e(u)\, du \quad (۴)$$

ملاحظه می‌شود که $\boldsymbol{V}$ عضوی از کلاس تبدیلات انتگرالی یعنی $\int \hat{K}(t,u)e(u)\, du$ می‌باشد که در آن $\hat{K}(t,u) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)x_i(u)$. تابع $\hat{K}(t,u)$ را هستهٔ تبدیل گویند.

۴) $\boldsymbol{W} : F \longrightarrow F$. نتیجه‌ی نگاشت بردار $\boldsymbol{f}$ از F بتوی E و برگشت آن برداری است که i اُمین عنصرش برابر است با:

$$f_i \xrightarrow{\boldsymbol{W}} \int_{0}^{T} x_i(t) \sum_{k=1}^{n} f_k x_k(t)\, dt. \qquad (۵)$$

۴ مقایسه‌ی تحلیل داده‌های تابعی و طولی

هر چند که تحلیل داده‌های تابعی و تحلیل داده‌های طولی[۱۴]، تا اندازه‌ای در اهداف و ساختار متفاوت هستند؛ امّا ویژگی‌های مشترکی نیز دارند. بر این اساس در این بخش به بررسی ارتباط این شاخه از آمار با تحلیل داده‌های تابعی می‌پردازیم. این بخش را با تعریف داده‌های طولی شروع می‌کنیم:

تعریف ۱ :(داده‌های طولی) مشاهده‌ی تعداد معینی از واحدها در طول زمان [۲].

با توجه به تعریف فوق، این داده‌ها را می‌توان از دو بُعد مورد توجه قرار داد: بُعد زمانی و بُعد مقطعی. برای توضیح بیشتر ابتدا به معرفی ساختار داده‌های طولی می‌پردازیم.

در تحلیل داده‌های طولی ، واحدهای مشخصی در زمان‌های متوالی مورد بررسی قرار گرفته و متغیر پاسخ اندازه‌گیری می‌شود. چنین داده‌هایی را می‌توان به صورت زیر نمایش داد:

جدول ۱. ساختار داده‌ها در مطالعات طولی

زمان \ واحد	۱	…	i	…	N
۱	y_{11}	…	y_{i1}	…	y_{N1}
$\vdots$	$\vdots$	…	$\vdots$	…	$\vdots$
t	y_{1t}	…	y_{it}	…	y_{Nt}
$\vdots$	$\vdots$	…	$\vdots$	…	$\vdots$
T	y_{1T}	…	y_{iT}	…	y_{NT}

در این جدول y_{it} مقدار متغیر پاسخ مربوط به واحد i اُم در زمان t اُم می‌باشد. T و N نیز به ترتیب معرّف تعداد دوره‌های زمانی و تعداد واحدهای تحت مطالعه می‌باشند. سطر (زمان)های این جدول، ماهیت مقطعی داده‌ها و ستون (واحد)های آن، ماهیت زمانی داده‌ها را نشان می‌دهند. به عنوان مثال، اگر داده‌هایی را که در زمان t اُم جمع آوری شده‌اند در نظر بگیریم در واقع واحدهای اول تا N اُم را تنها در یک مقطع از زمان مورد بررسی قرار داده‌ایم. بنابراین برای هر دوره‌ی زمانی، داده‌ها مقطعی می‌باشند. امّا اگر به طور مثال داده‌های مربوط به واحد i اُم را در نظر بگیریم تنها یک واحد را در زمان‌های متوالی مورد بررسی قرار داده‌ایم، بنابراین داده‌ها اثر زمان را نشان می‌دهند. در نتیجه، برای توصیف داده‌های طولی باید از مدل‌هایی استفاده شود که در آن ماهیت مقطعی و زمانی این گونه داده‌ها لحاظ شود.

[۱۴] Longitudinal Data Analysis(LDA)

داده‌های تابعی نیز مانند داده‌های طولی با اندازه‌گیری‌های مکرر روی هر آزمودنی و در طول زمان بدست می‌آیند. امّا زمان‌های جمع آوری در داده‌های تابعی به هم فشرده و نزدیک می‌باشند و این یکی از موارد تفاوت این دو می‌باشد. در تحلیل داده‌های تابعی ، اغلب اندازه‌گیری‌ها با دستگاه‌های حسگر خودکار و بازه‌ی زمانی فشرده انجام می‌گیرد. امّا در تحلیل داده‌های طولی ، اندازه‌گیری‌ها معمولاً پراکنده می‌باشند و با فواصل زمانی نامنظم از آزمودنی‌ها جمع آوری می‌شوند. از طرف دیگر، موضوع علمی این دو گرایش متفاوت می‌باشد: تحلیل داده‌های تابعی نگرش کاوشگرانه[۱۵] و تحلیل داده‌های طولی نگرش تأییدی[۱۶] دارد؛ به عبارت دیگر در تحلیل داده‌های تابعی با نمایش و توصیف داده‌ها برای تأکید بر مشخصه‌های مورد بررسی و شاید ورودی برای تحلیل بیشتر مواجه هستیم در حالی که در تحلیل داده‌های طولی بیشتر روی روش‌های استنباطی متمرکز است. این تفاوت را می‌توان در استفاده از توابع همبستگی-زمانی (خود همبستگی) مشاهده نمود؛ توابع همبستگی که در متون تحلیل داده‌های تابعی مورد استفاده قرار گرفته‌اند از یک نگاه توصیفی برای نمایش وابستگی‌های زمانی در منحنی‌ها برخوردار می‌باشد درحالی که برآورد چنین همبستگی‌هایی یکی از اهداف اصلی تحلیل داده‌های طولی و به منظور استخراج تحلیل معتبر بوده است. موضوع دیگری که به رویکردهای متفاوتی در دو گرایش منجر شده است، وجود داده‌های گمشده می‌باشد؛ این موضوع در تحلیل داده‌های تابعی دارای اهمیت کمی می‌باشد ولی در تحلیل داده‌های طولی به دلیل ساختار متفاوت، جستجو و استفاده از روش‌های دقیق‌تر برای برآورد داده‌های گمشده مورد نیاز است [۷].

علیرغم تفاوت‌های فوق‌الذکر، دو گرایش، اهداف مشترک زیادی دارند که در میان آن‌ها می‌توان به دو مورد زیر اشاره نمود: اول این‌که، در هر دو می‌توان از ابزارهای توصیفی نظیر میانگین به‌عنوان خلاصه‌ای از مشاهدات، استفاده نمود. دوم اینکه، تشخیص و استخراج الگوهای موجود در تغییرات بین منحنی‌ها و در هر دو گرایش شباهت دارند؛ این بدان معنا می‌باشد که چون در هر دو، واحدها در طول زمان مشاهده و بررسی می‌شوند می‌توان تغییرات صفت مورد اندازه‌گیری روی زمان را مورد بررسی قرار داد و به راحتی آزمودنی‌ای را که نسبت به بقیه متمایز هستند تشخیص داد. برای جزئیات بیشتر در این موضوع به [۷] مراجعه کنید.

۵ آماره‌های توصیفی برای داده‌های تابعی

آمار توصیفی برای تحلیل داده‌های تابعی را با معرفی آماره‌های توصیفی مکانی و پراکندگی بررسی می‌کنیم.

۱.۵ میانگین و واریانس تابعی

تابع میانگین با معدل‌گیری نقطه به نقطه محاسبه می‌شود. از این رو، تابع میانگین به صورت متوسط

[۱۵] Exploratory
[۱۶] Confirmatory

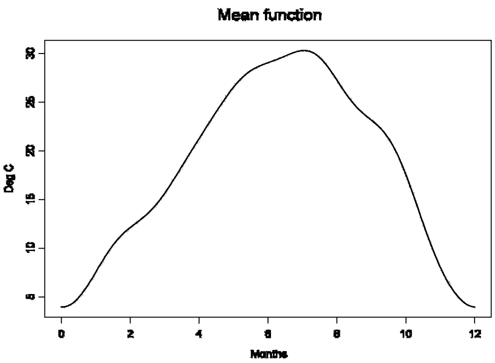
تکرارهای عرضی توابع مشاهدات تعریف می‌شود:

$$\bar{x}(t) = \frac{۱}{n}\sum_{i=۱}^{n} x_i(t). \qquad (۶)$$

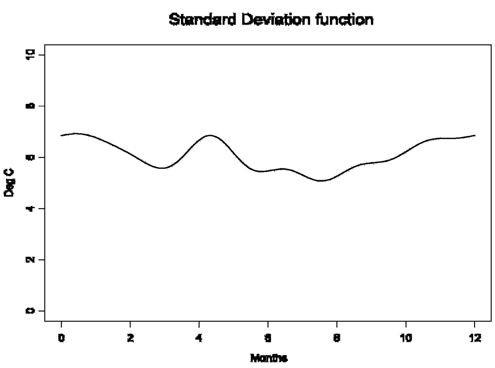
حال تابع واریانس نمونه‌ای برای هر نقطه t، $t \in \mathrm{I} = [۰, T]$، مشابه با واریانس نمونه‌ای و به صورت

$$Var_x(t) = \frac{۱}{n-۱}\sum_{i=۱}^{n}[x_i(t) - \bar{x}(t)]^۲, \qquad (۷)$$

تعریف می‌شود. در واقع این تابع، مقدار واریانس را در هر مقدار t محاسبه می‌کند.



شکل ۲. برآورد تابع میانگین دمای ایران در سال ۲۰۰۶. این منحنی با استفاده از داده‌های دما مربوط به ۱۰۲ ایستگاه هواشناسی محاسبه شده است.



شکل ۳. برآورد تابع واریانس دمای ایران در سال ۲۰۰۶. این منحنی با استفاده از داده‌های دما مربوط به ۱۰۲ ایستگاه هواشناسی محاسبه شده است.

۲.۵ تابع کوواریانس و همبستگی

در تحلیل چندمتغیره، میزان همبستگی بین هر جفت از متغیرها به وسیله‌ی مقدار کوواریانس نشان داده می‌شود؛ مشابه با آن، چنین عملی در حالت تابعی با ثابت نگه‌داشتن دو مقدار زمانی یعنی $t_۱$ و $t_۲$ و از طریق محاسبه‌ی

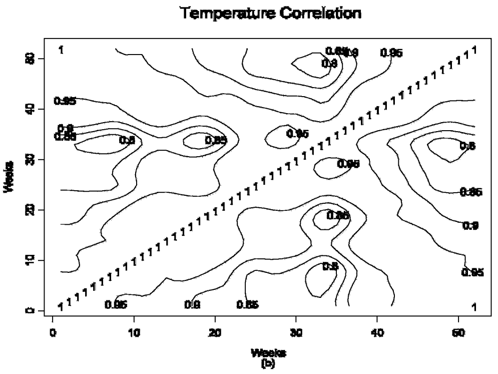
$$Cov_x(t_۱, t_۲) = \frac{۱}{n-۱}\sum_{i=۱}^{n}\{x_i(t_۱) - \bar{x}(t_۱)\}\{x_i(t_۲) - \bar{x}(t_۲)\}, \qquad (۸)$$

امکان‌پذیر می‌باشد. با توجه به رابطه‌ی فوق، تابع همبستگی[17] متناظر برابر است با:

$$Corr_x(t_۱, t_۲) = \frac{Cov_x(t_۱, t_۲)}{\sqrt{Var_x(t_۱)Var_x(t_۲)}}. \qquad (۹)$$

همان طور که مشاهده می‌کنید فرمول (۹) نیز با فرمول ضریب همبستگی برای هر جفت از داده‌های چندمتغیره مشابهت دارند. امّا براساس آن چه در بالا دیده شد در تحلیل داده‌های تابعی برای بررسی تغییرات به جای ماتریس‌ها از عملگرها استفاده می‌شوند. از این رو، می‌توان برای نمایش این ناحیه از روش ترسیمی و از نمودارهای خطوط تراز[18] یا سه بُعدی بهره برد.

[17] Correlation function
[18] Contour plots



شکل ۴. نمایش نمودار خطوط تراز تابع همبستگی برای دما در سال ۲۰۰۶. این نمودار با استفاده از داده‌های ۱۰۲ ایستگاه هواشناسی محاسبه شده است.

۳.۵ توابع کواریانس و همبستگی متقاطع

در بررسی پدیده‌های تابعی ممکن است دو یا چند متغیر تحت بررسی قرار گیرند. مثلاً در تحلیل داده‌های هواشناسی ممکن است متغیرهای دما و مقدار بارندگی مورد مطالعه قرار گیرند. در این گونه موارد، امکان بررسی ساختار همبستگی این دو متغیر به یکدیگر با استفاده از توابع کوواریانس یا همبستگی متقاطع وجود دارد.

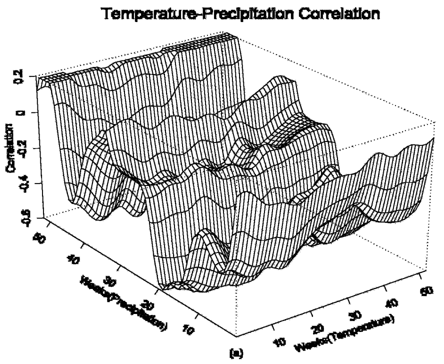
اگر مشاهدات ما به صورت زوج‌های $(x_i(t), y_i(t))$ باشند؛ آن‌گاه تابع کواریانس متقاطع[۱۹] این دو متغیر به صورت:

$$Cov_{X,Y}(t_1, t_2)=\frac{1}{n-1}\sum_{i=1}^{n}\{x_i(t_1)-\bar{x}(t_1)\}\{y_i(t_2)-\bar{y}(t_2)\},$$

و تابع همبستگی‌متقاطع[۲۰] آن‌ها به شکل:

$$Corr_{X,Y}(t_1, t_2)=\frac{Cov_{X,Y}(t_1, t_2)}{\sqrt{Var_X(t_1)Var_Y(t_2)}},$$

تعریف شود.



شکل ۵. نمایش سه بُعدی تابع همبستگی متقاطع بین دما و میزان بارندگی در سال ۲۰۰۶. این نمودار با استفاده از داده‌های ۱۰۲ ایستگاه هواشناسی محاسبه شده است.

در این مورد نیز، امکان نمایش بصری چنین توابعی با استفاده از نمودارهای خطوط تراز و یا سه بُعدی وجود دارد. برای جزئیات بیشتر به [۱] مراجعه شود.

۶ تحلیل مؤلفه‌های اصلی تابعی

همان‌طور که یکی از راهکارهای مناسب برای تلخیص داده‌های چندمتغیره استفاده از تحلیل مؤلفه‌های اصلی[۲۱] می‌باشد. این ابزار در حالت تحلیل داده‌های تابعی نیز برای تلخیص داده مورد استفاده قرار می‌گیرد. تحلیل مؤلفه‌های اصلی تابعی[۲۲] برای تشخیص ساختار داده‌ها و استخراج الگوهای مهم پراکندگی موجود در بین داده‌ها به‌کار می‌رود.

از دیدگاه آماری، در تحلیل مؤلفه‌های اصلی در پی یافتن مجموعه‌ای از ترکیبات خطی ناهمبسته‌ایم که تمام

[۱۹] Cross-covariance function
[۲۰] Cross-correlation function
[۲۱] Principal Components Analysis(PCA)
[۲۲] Functional Principal Components Analysis(FPCA)

تغییرات موجود در مجموعه داده‌ها را شرح دهند. به این ترکیبات خطی ناهمبسته، مؤلفه‌های اصلی [۲۳] گویند و از جایگذاری مقادیر مشاهده شده برای این ترکیبات خطی، امتیاز یا نمره آن مشاهده نامیده می‌شود. به دلیل آن‌که تعداد کمی از مؤلفه‌های اصلی می‌توانند اکثر تغییرات موجود در متغیرهای اولیه را بیان کنند؛ استفاده از این تعداد به جای استفاده از p متغیر اولیه مناسب خواهد بود.

به طور کلی مراحل استخراج مؤلفه‌های اصلی در یک آنالیز چند متغیره را می‌توان به‌صورت زیر بیان نمود: ۱− ترکیب خطی $z_1 = \alpha'_1 x$ (امتیاز یا نمره) را با بیشترین واریانس ممکن پیدا می‌کنیم، ۲− ترکیب خطی دیگری از متغیرها یعنی $z_2 = \alpha'_2 x$ را با بیشترین واریانس ممکن و ناهمبسته با z_1 را پیدا می‌کنیم، ۳− این فرآیند را تا استخراج حداکثر p مؤلفه اصلی ادامه می‌دهیم. امّا در عمل معمولاً عمل استخراج تا یافتن فقط m مؤلفه اصلی مورد توجه می‌باشد به طوری که این تعداد مؤلفه‌های اصلی بیشترین تغییرات موجود در متغیرهای اولیه را شامل شود. لازم به ذکر است که در هر مرحله قیود نرمالیدن مناسب یعنی $\alpha'_j\alpha_j = 1$، $j = 1, 2, \cdots, m$، لحاظ می‌شود تا فرآیند استخراج فوق معتبر باشد. بعلاوه قید ناهمبسته بودن مؤلفه‌ها ایجاب می‌کند که مؤلفه اصلی i اُم نسبت به تمامی مؤلفه‌های اصلی قبلی در رابطه $\alpha'_i\alpha_j = 0$، $j = 1, 2, \cdots, i-1$، صدق کند [۲]. مشابه با فرآیند بالا فرآیند استخراج مؤلفه‌ی اصلی تابعی تعریف می‌شود که در ادامه به آن اشاره می‌شود. همان طور که مشاهده خواهید کرد با انجام تغییراتی در رویکرد چندمتغیره بالا رویکرد تابعی استخراج مؤلفه‌های اصلی تابعی تعریف می‌شود. توجه شود که در حالت تابعی، به دلیل آن‌که مشاهدات به صورت تابع هستند هر یک از مؤلفه‌ی اصلی با تابع وزنی $\psi(t)$ (به جای بردار وزنی α) مشخص و روی بازه‌ی زمانی متناظر با داده‌های تابعی تعریف می‌شود.

فرض کنید داده‌های تابعی $x_1(t), \cdots, x_n(t)$ موجود باشند؛ آنگاه مشابه با حالت چند متغیره رویه‌ی زیر را برای استخراج مؤلفه‌های اصلی تابعی دنبال می‌کنیم:

۱− تابع وزنی $\psi_1(t)$ با بیشترین واریانس امتیازات ممکن یعنی $\frac{1}{n}\sum_{i=1}^{n}(z_{i1} - \bar{z})^2$ پیدا می‌کنیم به طوری که $z_{i1} = \int \psi_1(t)x_i(t)dt$،

۲− یافتن تابع مؤلفه‌ی اصلی دوم $\psi_2(t)$ را با بیشترین واریانس امتیازات مورد توجه قرار می‌دهیم به طوری که این تابع ناهمبسته با $\psi_1(t)$ باشند یعنی $\int \psi_2(t)\psi_1(t)dt = 0$،

۳− مراحل ۱ و ۲ را تا یافتن تعداد مطلوب از مؤلفه‌های اصلی تابعی ادامه می‌دهیم. مشابه با حالت چندمتغیره، برای حصول جواب رویه‌ی فوق باید قیود نرمالیدن یعنی دارا بودن طول واحد برای هر مؤلفه برقرار باشد؛ به عبارت دیگر برای مؤلفه‌ی اصلی $\psi(t)$ باید داشته باشیم $||\psi||_2 = \int \psi(t)^2 dt = 1$. بعلاوه شرط متعامد مؤلفه‌های اصلی تابعی نیز باید با توجه به تعریف اولیه در فرآیند استخراج لحاظ شود. به عبارت دیگر، تابع مؤلفه‌ی اصلی تابعی i اُم و تمام مؤلفه‌های اصلی تابعی‌های قبلی در رابطه‌ی $\int \psi_i(t)\psi_j(t)dt = 0$، $j = 1, 2, \cdots, i-1$، صدق می‌کنند.

---
[۲۳] Principal Components (PC)

به طور کلی با تعریف تابع کوواریانس نمونه‌ای به صورت

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

عملگر کوریانس نمونه‌ای برابر

$$(\widehat{K}\widehat{\psi})(v)=\int \widehat{K}(u,v)\widehat{\psi}(u)du,$$

خواهد بود. این عملگر از L_2 به L_2 تعریف می‌شود و متقارن و نیمه معین مثبت می‌باشد. اگر $\widehat{\theta}_j$ و $\widehat{\psi}_j(.)$ به ترتیب ویژه مقدار و ویژه تابع j اُم عملگر کواریانس نمونه باشند؛ آنگاه این دو کمیت در معادله‌ی ویژه‌ی

$$(\widehat{K}\widehat{\psi}_j)(u)=\widehat{\theta}_j\widehat{\psi}_j(u), \tag{10}$$

صدق می‌کنند.

رمسی و سیلورمن [۶] برای محاسبه‌ی مؤلفه‌های اصلی تابعی سه روش را ذکر کرده‌اند که هر یک از آنها جواب‌های تقریبی را برای معادله ویژه (۱۰) نتیجه می‌دهند. در روش اول، که به روش گسسته‌سازی معروف است؛ معمولاً هر آزمودنی در فواصل زمانی برابر در دامنه تعریف توابع ارزیابی می‌شود. در این روش ماتریس مشاهدات را با استفاده از تحلیل مؤلفه‌های اصلی معمولی تحلیل می‌کنیم و مقادیر ویژه و بردارها را به‌دست می‌آوریم؛ آن‌گاه برای تبدیل بردارهای ویژه به توابع ویژه، لازم است این بردارها را به بردارهایی به طول واحد تبدیل و سپس آنها را با هموارسازی مناسب درونیابی کنیم.

در رویکرد دوم فرض می‌شود که توابع $x_i(t)$ را می‌توان بر حسب یک مجموعه از G تابع پایه بیان کرد. از این رو، اگر $\phi_1(t),\ldots,\phi_G(t)$ توابع پایه مورد استفاده باشند، آن‌گاه

$$x_i(t)=\sum_{g=1}^{G}c_{ig}\phi_g(t).$$

به طور مشابه با نوشتن هر تابع ویژه بر حسب توابع پایه معادله ویژه به راحتی قابل حل است.

روش سوم، بکارگیری طرح مربع‌بندی عددی[24] برای تقریب انتگرال موجود در معادله ویژه را مورد توجه قرار می‌دهد. برای جزئیات بیشتر در مورد این روش‌ها می‌توان به [۶] مراجعه نمود.

لازم به ذکر است که در این مقاله از روش‌ها دوم برای حل معادلات استفاده شده است.

۷ تفسیر مؤلفه‌های اصلی در داده‌های تابعی

همان‌طور که می‌دانیم، اصولاً تفسیر نتایج حاصل از تحلیل مؤلفه‌های اصلی در بسیاری از مسایل آنالیز چند متغیره کار آسانی نیست. در داده‌های تابعی، مشکل دو چندان می‌شود، زیرا علاوه بر مشخصه‌ی تصادفی بودن مشاهدات (منحنی‌ها)، هر مشاهده تابعی از یک متغیر دیگر مانند t (زمان) است. بنابراین، همان طور که در بالا توضیح داده‌ایم بردارهای ویژه ، به توابع ویژه تبدیل می‌شوند و در نتیجه بر خلاف آنالیز چند متغیره، z_j ها یا نمره‌ی هر مشاهده از جنس آن مشاهده نمی‌باشد. لذا در بعضی از موارد، ابزارها برای تفسیر مؤلفه‌های اصلی در داده‌های تابعی متفاوت از ابزارهای موجود برای این منظور در داده‌های چند متغیره است [۶].

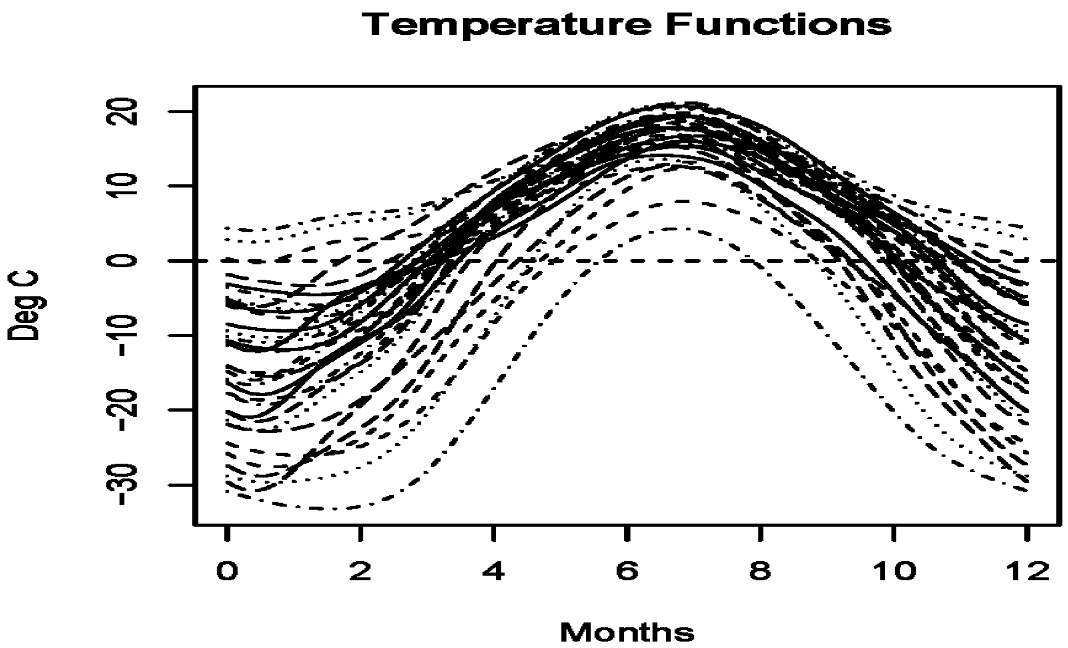
[24] Numerical quadrature scheme

یک روش مفید برای تفسیر مؤلفه‌های اصلی تابعی، بررسی نمودار تابع میانگین و توابع بدست آمده از اضافه یا کم کردن یک ضریب مناسب از تابع مؤلفه‌ی اصلی مورد سوال است. شکل (۸) چنین نموداری را برای داده‌های دما نشان می‌دهد. در این حالت، منحنی پیوسته، تابع میانگین دما، و منحنی‌های $(+)$ و $(-)$، اثرات اضافه یا کم کردن ضریبی از توابع ویژه را نشان می‌دهد. این روش به طور قابل توجه‌ای اثرات اولین دو مؤلفه‌ی اصلی را نمایان می‌کند. در ساختن این نمودار، لازم است تا ضریبی مناسب از توابع مولفه اصلی انتخاب شود. ثابت c را به صورت

$$c^{2} = T^{-1}\,\|\hat{\mu}-\bar{\mu}\|^{2} = T^{-1}\int_{I}(\hat{\mu}(t)-\bar{\mu})^{2}\,dt,$$

تعریف می‌کنیم که در آن T طول بازه I و $\hat{\mu}(t)=\frac{1}{n}\sum_{i=1}^{n}x_i(t)$ و $\bar{\mu}=T^{-1}\int_I \hat{\mu}(t)\,dt$. سپس نمودار $\hat{\mu}(.)$ و $\hat{\mu}(.)\pm c\hat{\psi}_i(.)$ با هم رسم می‌شوند تا از مقایسه‌ی این دو با هم نتایج قابل تفسیری بدست آید. لازم بذکر است که انتخاب c به رفتار $\hat{\mu}(.)$ نیز بستگی دارد.

به عنوان یک مثال کاربردی، یک مجموعه داده واقعی مورد تحلیل قرار گرفته است. داده‌های واقعی شامل دمای ثبت شده در ۳۵ ایستگاه هواشناسی متفاوت در کانادا برای سال ۱۹۸۲ است (شکل (۶)). هدف، استخراج عوامل و الگوهای تغییرپذیری دماست که بوسیله‌ی تحلیل مولفه‌های اصلی صورت پذیرفته است. همچنان که شکل (۷) نشان می‌دهد مقادیر ویژه اول یعنی $\hat{\theta}_1$ نشان دهنده‌ی آن است که عمده‌ترین عامل تغییرپذیری در دما، مربوط به اولین مؤلفه‌ی اصلی است. علاوه بر آن، اولین چهار مؤلفه‌ی اصلی، بیش از ۹۹ درصد از کل تغییرپذیری را به خود اختصاص می‌دهند که سهم هر کدام به‌طور جداگانه به ترتیب ۸۹/۳، ۸/۳، ۱/۶، ۰/۵ درصد است.



شکل ۶. دمای کانادا در سال ۱۹۸۲ که از ۳۵ ایستگاه هواشناسی مختلف در سراسر کانادا جمع‌آوری شده‌اند.

سهم بقیه‌ی عوامل رویهم کمتر از ۰/۵ درصد می‌باشد. در نتیجه، تنها اولین چهار عامل تغییرپذیری را مورد توجه قرار می‌دهیم.

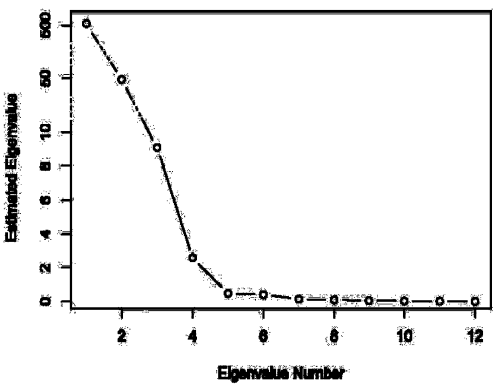
برای بررسی و تفسیر مؤلفه‌های اصلی، در اینجا $c = ۰/۲$ انتخاب شده است. این انتخاب تنها به این دلیل بوده است تا نتایج به‌طور واضح قابل رویت باشند.

بررسی دو شکل (۸) و (۹) و $\widehat{z}_{ij}$ ها موارد زیر را آشکار می‌کند:

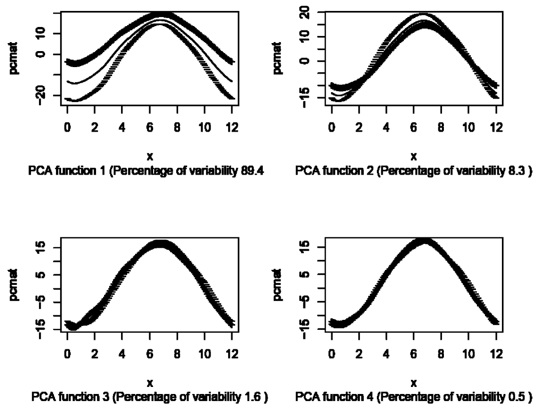
اولین مؤلفه اصلی برآورد شده نشان می‌دهد اکثر تغییرپذیری (۸۹ درصد) در داده‌ها از تفاوت‌هایی که بین دما در تابستان و زمستان وجود دارد ناشی می‌شود. دومین مؤلفه‌ی اصلی برآورد شده، تغییرپذیری دما وقتی که از زمستان به تابستان حرکت می‌کنیم را نشان می‌دهد. به عبارت دیگر، این مؤلفه، تغییرپذیری از متوسط تفاوت بین دما در زمستان و دما در تابستان را منعکس می‌کنند. این مولفه، برای زمستان به طور مثبت و برای تابستان به طور منفی تاثیرگذار است. بنابراین، این مولفه به مناطقی که اختلاف دمای زمستان و تابستان در آنها کوچک است یک نمره مثبت بالا اختصاص می‌دهد. در مقابل، نمرات منفی بزرگ به مناطقی که در آنها تابستان گرم نیست و زمستان سرد است اختصاص می‌یابد. سومین مؤلفه‌ی اصلی متناظر با اثر یک جا به جایی زمانی است که همراه با کمی افزایش در هم دما و هم فاصله‌ی بین زمستان و تابستان می‌باشد. چهارمین مؤلفه‌ی اصلی نیز مربوط به اثری است که سبب می‌شود که بهار دیرتر آغاز شود و پاییز زودتر به پایان برسد.

۸ بحث و نتیجه‌گیری

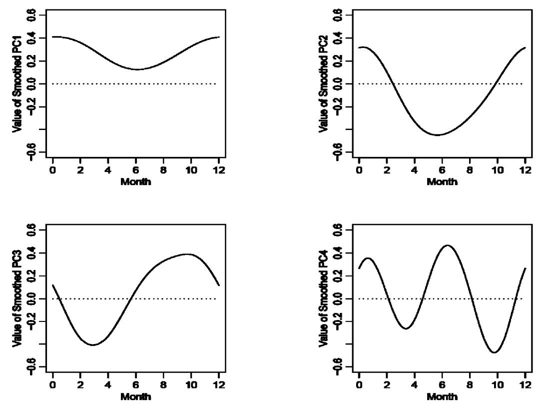
در بسیاری از موارد داده‌های مشاهده شده اولیه در مورد پدیده‌هایی که ذاتاً دارای ماهیت تابعی می‌باشند گسسته هستند. در چنین شرایطی باید ابتدا به این داده‌های گسسته، منحنی‌های پیوسته برازش داد و سپس از روش‌های تابعی برای تحلیل آنها استفاده نمود. از آن جا که داده‌های تابعی در واقع تعمیمی از داده‌های چندمتغیره می‌باشند؛ به مقایسه این داده‌ها با داده‌های تابعی پرداختیم و همان طور که مشاهده شد، در بسیاری



شکل ۷. ویژه مقدارهای برآورد شده برای داده‌های دمای کانادا در سال ۱۹۸۲.



شکل ۸. تابع میانگین دما و اثرات اضافه (+) و کم کردن (−) ضریب مناسبی از اولین چهار تابع PC.



شکل ۹. چهار مؤلفه‌ی اصلی دما بر داده‌های دمای کانادا.

از موارد مرحله‌ی گذار از حالت چندمتغیره به تابعی شامل تغییر یک علامت مجموع به انتگرال می‌باشد [۳]. همچنین یادآور می‌شویم که اگر داده‌های طولی بر روی یک بازه‌ی زمانی فشرده جمع‌آوری شوند می‌توان آنها را به صورت تابعی در نظر گرفت. آماره‌های توصیفی برای داده‌های تابعی نیز به صورت تعمیمی از این آماره‌ها در حالت چندمتغیره معرفی و به عنوان مثال، آماره‌های توصیفی برای داده‌های دما و بارندگی ایران محاسبه شدند. پس از آن، بدنبال معرفی تحلیل مؤلفه‌های اصلی برای دادهای چندمتغیره؛ تحلیل مؤلفه‌های اصلی تابعی را معرفی کردیم و در انتها نیز، با استفاده از داده‌های واقعی، چگونگی محاسبه و تفسیر مؤلفه‌های اصلی برای داده‌ها دما توضیح داده شد.

تشکر و قدردانی

از دو داور محترم که با پیشنهادات ارزنده‌ی خود ما را در نگارش بهتر مطالب یاری نموده‌اند کمال تشکر را داریم.

مراجع

[۱] تازیکه‌میاندره، ن. و حسینی‌نسب، م. ا. (۱۳۸۷)، تحلیل تابعی دما و بارندگی در ایران با استفاده از مؤلفه‌های اصلی تابعی، پذیرفته شده در مجله پژوهش‌های آماری ایران.

[2] Diggle, P., Heagerty, P., Liang, K. L. and Zeger, S. (2002), *Analysis of Longitudinal Data*, 2nd Ed., edition Oxford University Press, USA.

[3] Jolliffe, I. T. (2002), *Principal Component Analysis*, 2nd Ed., Springer, New York.

[4] Ramsay, J. O. (1982), When the Data are Functions, *Psychometrika*, 47(4), 379-96.

[5] Ramsay, J. O., Silverman, B. W. (2002), *Applied Functional Data Analysis: Methods and Case Studies*, Springer, New York.

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

[7] Rice, J. A. (2004), Functional and Longitudinal Data Analysis: Perspectives on Smoothing, *Statistica Sinica*, 14, 631-647.