

نمایش جنبه های مهم تغییرات مجموعه بزرگی از منحنیهای مشابه

ام.سی. جونز جی.ا. رایس

مترجم: آیت الله موسوی^۱

چکیده

نمایش تعداد زیادی از منحنیها بر روی یک نمودار (به طوری که هر منحنی بر روی سایر منحنیها قرار گیرد یا در قسمتهای زیادی با یکدیگر تلاقی داشته باشند) از نظر ظاهری ناخوشایند و از جنبه اطلاع دهی ضعیف است. در این مقاله با استفاده از روش تحلیل مؤلفه های اصلی، تعداد محدودی از منحنیها طوری انتخاب می شوند که نحوه تغییرات موجود بین همه منحنیها را تا حد زیادی بیان می کنند.

۱. طرح مسأله

شکل ۱ ظاهر زیبایی ندارد. در این شکل مجموعه ای از ۱۰۰ منحنی مشابه با یک روش ثابت و با استفاده از ۱۰۰ مجموعه داده شبیه سازی شده، روی یک نمودار رسم شده اند. روشن است که برای نمایش جنبه های مهم تغییرات این مجموعه منحنی، شکل ۱ از چند جهت با شکست مواجه می شود. رفتار هر یک از منحنیها در اکثر قسمتهای نمودار، در انبوه حاصل از سیاهی ایجاد شده توسط سایر منحنیها گم می شود. هر چه تعداد این منحنیها افزایش یابد، نمودار حاصل به جای این که اطلاع دهی بیشتری داشته باشد، گیج کننده تر می شود. در بخش ۲ روش ساده ای برای نمایش قسمت عمده اطلاعات پنهان شده در شکل ۱ ارائه می شود. برای انجام این کار، از روش تحلیل مؤلفه های اصلی استاندارد استفاده می کنیم. نتیجه این کار، مجموعه ای از دو یا سه شکل است که هر یک تعداد محدودی، مثلاً ۳، منحنی را در بردارند. لازم به تذکر است که در این روش فقط زیرمجموعه ای از منحنیهای اولیه

نمایش داده می شوند، نه اینکه اطلاعات موجود در منحنیهای اولیه ترکیب شوند تا تسوابع جدیدی به دست آیند. در عوض تلاش می شود تا منحنیهایی را انتخاب کنیم که جنبه های مهم تغییرات مجموعه منحنیها را به بهترین وجه نمایش دهند. برای تشریح این روش، داده های شکل ۱ را به کار می بریم. در حقیقت هر یک از این ۱۰۰ منحنی، برآورد یک تابع چگالی به صورت زیر است: [۱]

$$\hat{f}(x) = 1/0.2 \sum_{i=1}^{100} h^{-1} \phi\{h^{-1}(x - X_i)\}$$

که در آن $X_1, X_2, \dots, X_{100}$ نمونه ای از توزیع نرمال استاندارد با تابع چگالی ϕ است. پارامتر h را پهنای پنجره یا پهنای نوار^۲ می نامند که یک پارامتر هموار کننده است و میزان هموار کردن داده ها را کنترل می کند. اخیراً اکثر کارهای انجام شده در این زمینه بر نحوه انتخاب h متمرکز بوده است. این مثال از جمله مواردی است که به کار بردن روش این مقاله برای آن می تواند مفید باشد. کاربرد دیگر این روش در داده های حاصل از

^۱ دانشگاه پیام نور، بروجرد

^۲ Band Width

$$Y_1 = e_1'X \quad \text{مؤلفه اصلی اول}$$

$$Y_2 = e_2'X \quad \text{مؤلفه اصلی دوم}$$

⋮

$$Y_k = e_k'X \quad \text{مؤلفه اصلی } k \text{ ام}$$

هر یک از این مؤلفه ها دارای N مشاهده است که آنها

را مرتب کرده و با $P_{(1)}, P_{(2)}, \dots, P_{(N)}$ نشان می دهیم. $P_{(1)}$

مینیمم، $P_{(N+1)}$ میانه، $P_{(N)}$ ماکسیمم این مشاهدات اند. حال

منحنیهای متناظر با این سه مشاهده را از بین منحنیها انتخاب و رسم می کنیم. معمولاً مؤلفه های اصلی اول و دوم، بیشترین تغییرات موجود در داده ها را توضیح می دهند. شکلهای ۲ و ۳، نتیجه کاربرد این روش، درباره منحنیهای شکل ۱ هستند. شکل ۲ مربوط به مؤلفه اصلی اول است که ۴۰٪ درصد از کل تغییرات را بیان می کند. منحنی خط چین کشیده تر مربوط به $P_{(1)}$ یک چگالی با پراکندگی کم است، منحنی خط چین عریضتر، یک چگالی با پراکندگی زیاد و منحنی پر رنگتر مربوط به میانه است که پراکندگی متوسطی دارد.

شکل ۳ مربوط به مؤلفه اصلی دوم است و نوع متفاوتی از

تغییر منحنیها را نمایش می دهد. مؤلفه اصلی دوم ۳۶٪ درصد از کل تغییرات را بیان می کند. این میزان تغییر با تغییراتی که توسط مؤلفه اول بیان می شود، همبستگی ندارد. در شکل ۳ منحنیهای مربوط به مینیمم و ماکسیمم، مشاهدات مؤلفه اصلی دوم (یعنی منحنیهای خط چین) در جهات مخالف یکدیگر، چوله هستند. منحنی مربوط به میانه، چولگی کمتری دارد (هرچند کاملاً متقارن نیست). این دو مؤلفه مجموعاً ۷۷ درصد از کل تغییرات را بیان می کنند، لذا مؤلفه اصلی سوم و بالاتر از آن از اهمیت کمتری برخوردارند و همچنین تغییراتی که توسط آنها بیان می شود به آسانی قابل تفسیر نیست.

در شکل ۲ قطعاً و در شکل ۳ تا حدی قابل توجه، تفسیر

جنبه های تغییر نمایش داده شده، از خود شکلها واضح است. اما همیشه این طور نیست. در چنین حالتی نمودارهای کمکی

اندازه گیری های مکرر^۳ یک صفت (بر روی انسان، حیوان و ...) در زمانهای مختلف یا تحت شرایط متفاوت است. در این مورد نیز مثال داده های سطح ازون شرح داده خواهد شد. منحنیهای رشد نیز از این قبیل اند.

۲. تحلیل مؤلفه های اصلی

از بین ۱۰۰ منحنی شکل ۱ باید تعداد محدودی از منحنیها را طوری انتخاب کنیم که جنبه های مهم تغییرات موجود در مجموعه منحنیها را به بهترین وجه منعکس کنند. یکی از روشهای مؤثر در انجام این کار، استفاده از تحلیل مؤلفه های اصلی است. در این حالت می توانیم هر یک از N تابع را به صورت برداری شامل k مقدار در نظر بگیریم. در منحنیهای رشد و طرح اندازه های تکرار شده نیز هر یک از N موجود (انسان، حیوان، قطعه و ...) تحت k وضعیت مختلف یا در k زمان متفاوت، مورد اندازه گیری قرار می گیرند. لذا یک ماتریس $N \times k$ از داده ها داریم که می توانیم تحلیل مؤلفه های اصلی استاندارد را نسبت به ماتریس کوواریانس مربوط به آنها به کار ببریم [۲]. فرض کنید

$$X'_{N \times k} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \dots & X_{Nk} \end{bmatrix}$$

اگر $S_{k \times k}$ ماتریس کوواریانس X' باشد و $e_1, \dots, e_k$

بردارهای ویژه ماتریس S باشند و

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$$

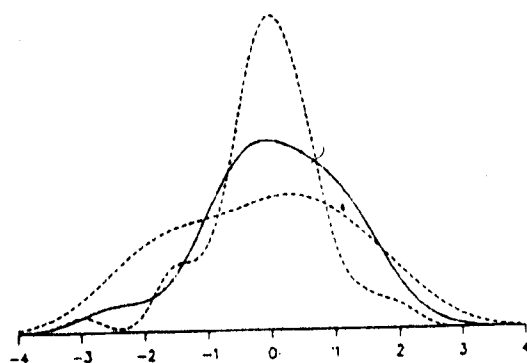
آنگاه مؤلفه های اصلی به ترتیب عبارت اند از:

^۳ Repeated Measures

یک از مؤلفه های اصلی اول و دوم، منحنیهای مربوط به $P_{(ii)}$ ، $P_{(i)}$ و $P_{(iv)}$ در شکلهای ۶ و ۷ رسم شده اند. خود مؤلفه های اصلی اول و دوم نیز در شکل ۸ نمایش داده شده اند. در اینجا رسم مؤلفه اصلی اول، در مقابل یک متغیر کمکی که در این مثال، شماره هفته در نظر گرفته شده، یک روند تغییرات را نشان می دهد (شکل ۹). در واقع قسمت عمیق نمودار مربوط به افزایش لایه های ازون در ماههای تابستان است. مؤلفه اصلی دوم یک تغییر خطی در ازون را در طول هفته نشان می دهد. در برخی هفته ها لایه ازون، افزایش، در برخی دیگر کاهش داشته و در اغلب هفته ها لایه ازون تقریباً ثابت بوده است.

توانایی این روش ساده در معنی دار تر کردن نمودارهای نامفهومی مانند شکلهای ۱ و ۵، شگفت انگیز است، و نشان می دهد که استفاده از روش تحلیل مؤلفه های اصلی که در اکثر نرم افزارهای رایانه ای در دسترس است، فواید عملی بسیاری دارد.

اینکه چه منحنیهایی باید در هر مسأله انتخاب شوند تا بیشترین تغییرات را بیان کنند، هنوز نیاز به کار بیشتری دارد. استفاده از مینیمم، ماکسیمم و میانه تنها یک پیشنهاد است که در این مقاله مورد استفاده قرار گرفت.

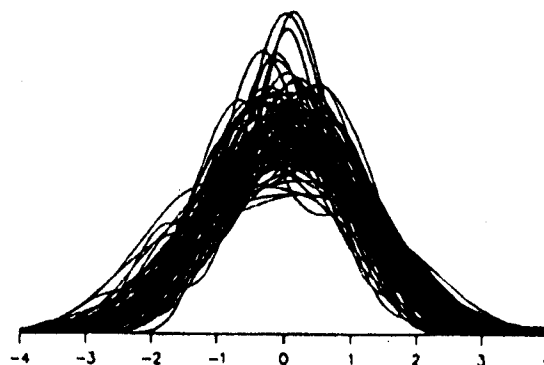


شکل ۲: منحنیهای حاصل از مؤلفه اصلی اول برای شکل ۱. منحنی بررنگتر میانه و منحنیهای خط چین، مینیمم و ماکسیمم هستند.

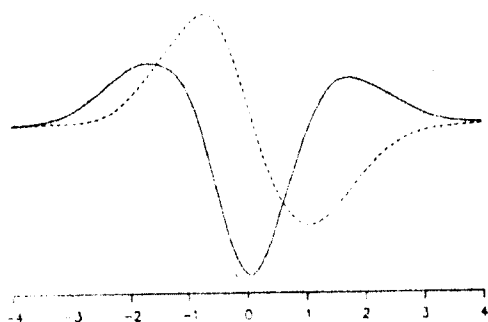
مؤلفه های اصلی، ممکن است مفید باشند. از این رو مؤلفه های اصلی متناظر با شکلهای ۲ و ۳، در شکل ۴ نمایش داده شده اند. برای تأیید ادعای ما در رابطه با پراکندگی و چولگی منحنیها، به تجمع اکثر منحنیهای مؤلفه اصلی اول در مرکز، نسبت به دو انتهای منحنیها، در شکل ۴ توجه کنید. قسمت عمیق مرکزی را با قله های کوچک دو طرف مقایسه کنید. به طور مشابه دومین مؤلفه اصلی تغییرات قابل ملاحظه برابر، ولی در خلاف جهت یکدیگر در شانه های توزیع را، نمایش می دهد. لذا در اینجا منحنیها از منحنی متوسط (تقریباً متقارن) تا منحنیهایی که به شانه های راست و چپ چوله اند، تغییر می کنند. نمودارهای کمکی دیگری نیز مانند نمودار جعبه ای مشاهدات مؤلفه های اصلی، قابل استفاده اند.

۳. مثال دوم: داده های لایه های ازون

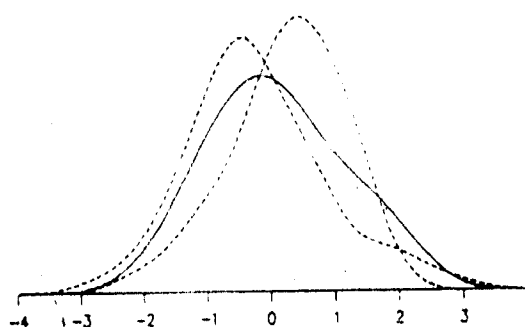
داده های مثال دوم مربوط به حداکثر لایه ازون است که برای $N = 47$ هفته از سال و در هر هفته $k = 7$ مقدار روزانه در منطقه معینی ثبت شده است [۳]. داده های خام در شکل ۵ نمایش داده شده اند و همانگونه که ملاحظه می شود تفسیر این شکل خیلی مشکل است. چهار مؤلفه اصلی اول این منحنیها به ترتیب ۶۱/۹، ۱۸/۳، ۷/۴ و ۵/۷ درصد از کل تغییرات را بیان می کنند. برای هر



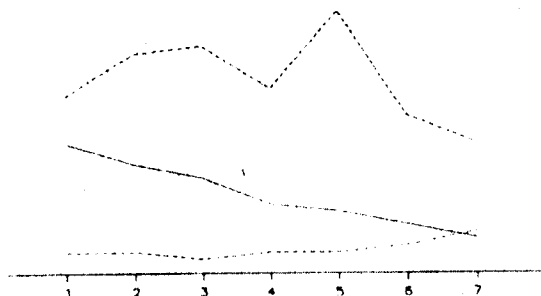
شکل ۱: مجموعه ای از برآوردهای تابع چگالی بر اساس ۱۰۰ نمونه مستقل به اندازه ۵۰ از توزیع نرمال استاندارد [۴]



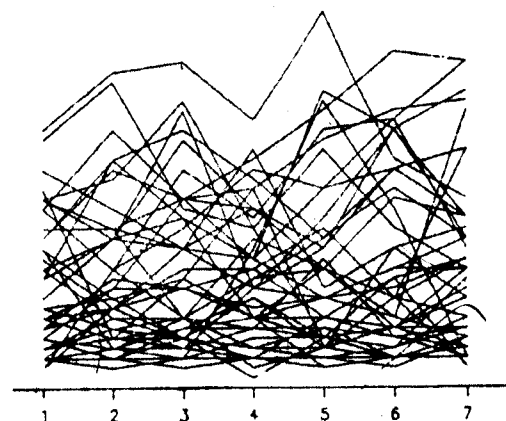
شکل ۴: مؤلفه های اصلی اول و دوم - مؤلفه اصلی اول (منحنی پررنگتر) و مؤلفه اصلی دوم (منحنی خط چین)



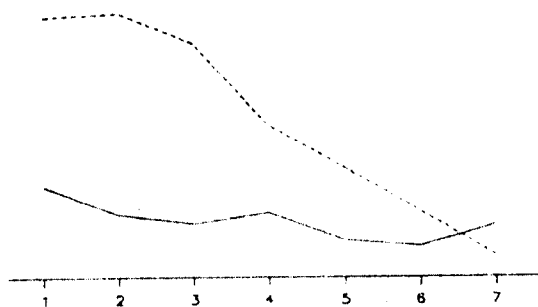
شکل ۳: منحنیهای حاصل از مؤلفه اصلی دوم برای داده های شکل ۱



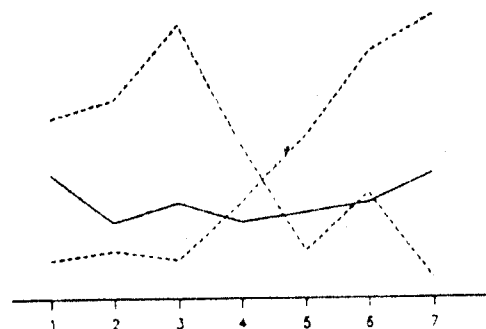
شکل ۶: منحنیهای حاصل از مؤلفه های اصلی اول برای داده های ازون - منحنی پررنگتر، میانه و منحنیهای خط چین مینیمم و ماکسیمم اند.



شکل ۵: داده های ازون



شکل ۸: مؤلفه های اصلی اول و دوم برای داده های ازون - مؤلفه اصلی اول (منحنی پررنگتر) و مؤلفه اصلی دوم (منحنی خط چین)



شکل ۷: منحنیهای حاصل از مؤلفه اصلی دوم برای داده های ازون