۱۳۸۰ /۱۱/ ۱۵

به نام خدا

برآوردیابی در مدلهای متغیر نهان
با استفاده از تحلیل مؤلفه های اصلی و تحلیل تناظر

به وسیله ی :

حمیدرضا خورشیدی

پایان نامه

ارائه شده به معاونت تحصیلات تکمیلی به عنوان بخشی از

فعالیتهای تحصیلی لازم برای اخذ درجه کارشناسی ارشد

در رشته ی:

آمارریاضی

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی و تصویب شده توسط کمیته پایان نامه با درجه :بسیار خوب

دکتر عبدالرسول برهانی حقیقی،استادیار بخش آمار (رئیس کمیته).....

دکتر جواد بهبودیان،استاد بخش آمار.....

دکترناهید سنجری فارسی پور،دانشیار بخش آمار.....

مهر ماه ۱۳۸۰

تقدیم به:

پدر و مادر مهربانم

سپاسگزاری

در ابتدای کلام ، پروردگاری که زیباییش را در آموختن، دیدن و شنیدن بر من نمایان ساخت سپاس می گویم.

و از استاد گرامی خود،دکتر عبدالرسول برهانی حقیقی که راهنمایی های ایشان مرا در تدوین این پایان نامه موفق ساخت،و استاد فرزانه دکتر جواد بهبودیان و خانم دکتر ناهید سنجری که در طول دوران تحصیل از آموخته های ایشان استفاده فراوان نمودم، بسیار قدر دانی می کنم.

همچنین از خانواده خود و دوست عزیزم،مجید عظیم محسنی،که همیشه در کنار من بوده اند تشکر فراوان دارم.

چکیده

برآوردیابی در مدلهای متغیر نهان

با استفاده از تحلیل مؤلفه های اصلی و تحلیل تناظر

به وسیله ی :

حمیدرضا خورشیدی

تحلیل مؤلفه های اصلی(PCA) و تحلیل تناظر (CA) اغلب در تحلیل داده های چندمتغیره بکارمیروند.در PCA، n مشاهده از m متغیر اولیه در یک ترکیب خطی به مشاهداتی با بعد کمتر تبدیل میشوند .این تبدیل بوسیله بردار ویژه های ماتریس کواریانس نمونه ای S صورت میگیرد. CA نیز معادل با PCA برای داده های طبقه ای با ویژگی کاهش بعد میباشد.با این تفاوت که در CA بردار ویژه های یک ماتریس وزنی از مشاهدات در تحلیل مورد استفاده قرار میگیرند.

در این پایان نامه بکار گیری بردار ویژه های اصلی،که حاوی بهترین اطلاعات ازمشاهدات هستند،برای برآورد پارامتر مدلهای متغیر نهان مد نظر میباشد.این پایان نامه شامل شش فصل میباشد.فصل اول مروری دارد بر تاریخچه.همچنین در این فصل نقش بردار ویژه ها در تجزیه مقدار منفرد(SVD) آمده است.

تعریف PCA و CA ، و ارتباط آنها با تجزیه مقدار منفرد ،در فصل دوم مورد بحث میباشد و بردار ویژه های مرتبط با SVD بعنوان شاخصهای سطری و ستونی PCA و CA تعریف میشوند.

در فصل سوم با تعریف کوتاهی از مدلهای خطی تعمیم یافته ،دو مدل از این کلاس معرفی میشوند که در آنها تابع پیش بینی کننده به صورت یک تابع خطی ویا درجه دوم از متغیر های نهان میباشد. پیش بینی کننده های خطی در این مدلها بصورتهای $\eta_{ij} = \alpha_j + \beta_j x_i$ و یا $\eta_{ij} = a_j - b_j(x_i - u_j)^2$ تعریف شده اند ،جاییکه x_i یک متغیر نهان و $i = 1,2,...,n; j = 1,2,...m$ میباشد و برای متغیر های پاسخ در این مدلها، توزیعهای نرمال، پواسن و یا برنولی مفروض است.

در فصل چهارم نشان داده میشود که PCA با برآورد MLE برای $\boldsymbol{\beta}$ در مدل خطی گاوسی معادل است ،و همچنین CA میتواند در تقریبِ برآورد MLE برای $\boldsymbol{u}$ بکار رود.

سازگاری حدی برآورد گرها ی پیشنهادی ، در فصل پنجم مورد ارزیابی قرار میگیرد و در انتها، در فصل ششم کاربرد بخشی از نتایج فوق برای داده های محیط زیستی بیان میگردد.

فهرست مطالب

فهرست جدول ها

عنوان **صفحه**

فهرست شکل ها

عنوان **صفحه**

مقدمه

فرض می‌کنیم برای n شی متمایز از یک جامعه m ویژگی (متغیر تصادفی) مورد نظر باشد. مشاهدات حاصل از m متغیر را برای هر کدام از n شی می‌توان به شکل آرایش ماتریسی نمایش داد بگونه‌ای که سطر i ام،شامل m ویژگی شی i ام می‌باشد،که آن را ماتریس مشاهدات مینامیم. هدف یک مطالعه بهترین استنباط از این ماتریس برای جامعهٔ آماری مورد نظر است. یک محقق ممکن است علاوه بر در دست داشتن مشاهدات فوق،یک سری فرضیات تئوری را نیز از تأثیر متغیرهای دیگر (متغیرهای کمکی) بر این مشاهدات (متغیرهای پاسخ) در نظر بگیرد. که منجربه استفاده از تئوری مدلها در مطالعهٔ او می‌گرددکه دراین تئوری کشف ارتباط بین متغیرهای کمکی و پاسخ بر اساس جمع‌آوری مشاهداتی از هردومتغیرصورت میگیرد. اگر به هر دلیل امکان مشاهدهٔ متغیر کمکی وجود نداشته باشد. این مدل در اصطلاح یک " مدل متغیرنهان" میباشد. که در آن متغیر کمکی به عنوان یک متغیر نهان (بی‌مشاهده) حضور دارد.

از جهت دیگر یک محقق ممکن است از تحلیل مولفه‌های اصلی (PCA) و تحلیل تناظر (CA) برای مشاهدات خود استفاده نماید. PCA و CA دو روش مرتبط با یکدیگر در کاهش بعد و توجیه ساختار داده‌های چند متغیره میباشند. در PCA ، اولین محورهای اصلی ماتریس کواریانس S بعنوان محورهایی که در یک بعد کمتر حداکثر بار اطلاعاتی موجود در S را به همراه دارند بعنوان مختصات جدید برای تصویر داده‌های اولیه بکار میروند. CA را نیز

معادل با PCA برای داده‌های طبقه‌ای میدانند. که در آن بجـای مـاتریس کواریـانس S اولیـن محورهای اصلی یک ماتریس وزنی از مشاهدات (که در فصل دوم به تعریـف آن مـی‌پردازیـم) بعنوان محورهای مختصات جدید مورد نظر میباشد. محورهای اصلی در اینجـا بـردار ویژه‌هـای ماتریسهای مذکور هستند.همانگونه که میدانیم در جبر ماتریسها مقادیر و بـردار ویژه‌هـای هـر ماتریس بیشترین نقش را در تجزیه‌های گوناگون آن دارند. از جملـهٔ ایـن تجزیـه‌ها میتـوان از تجزیه مقدار منفرد (SVD) نام برد،که تعمیم تجزیه طیفی ماتریسهای متقارن برای ماتریسهای غیر مربعـی اسـت و درآن هرماتریس $A_{n\times m}\equiv[a_{ij}]$ بـا رتبـهٔ k بـه حـاصلضرب دو مـاتریس متعامد $n\times k$ و $m\times k$ویک ماتریس قطری $k\times k$بصورت $A=P\Delta Q^{T}=\sum_{r=1}^{k}\lambda_r p_r q_r^{T}$ تجزیه میگردد، که مقادیر منفرد ماتریس A بترتیب نزولی $\lambda_1>\lambda_2>\ldots>\lambda_k$ در نظـر گرفتـه میشوند.برای آنکه نقش این تجزیه در جبر ماتریسها بیشتر نمایان شـود بـه قضیـه ای در ایـن ارتباط توجه میکنیم:

قضیه(۱-۱):برای هر ماتریس A با تجزیـه فـوق و هـر مقـدار صحیـح و مثبـت ρ کـه کـوچکـتر از k باشـد،خواهیـم داشـت: $\min_{B_{n\times m}:Rank(B)<\rho}\|A-B\|^2=\sum_{r=\rho+1}^{k}\lambda_r^2$ ، اگر و تنـها اگر $B=\sum_{r=1}^{\rho}\lambda_r p_r q_r^{T}$ باشد،که درآن نماد $\|.\|$ برابر نرم اقلیدسی است (Gabriel-1978).

بنابراین درمیان ماتریسهایی با رتبه کمتراز k نزدیکترین ماتریس بـه A در انـدازه اقلیدسـی،بـا مجموع اولین جملات در SVD ماتریس A برابـر اسـت و هـر انـدازه کـه اولیـن مقـادیرمنفرد ماتریس A که در این مجموع شرکت دارند در مقایسه با دیگر مقادیر بزرگ باشند،این نزدیکـی به ماتریس A بیشتر خواهد بود.نقش اولین بردارهای ویژه دراین قضیه، برای ماتریس کواریـاس نمونه‌ای S در تحلیل مؤلفه های اصلی و معادل آن برای ماتریس وزنی از مشاهدات در تحلیـل تناظر ظاهر میشود. که این موضوع در فصل دوم پایان‌نامه مورد بررسی قرارمیگردد.

هدف ما نشان دادن این نقش و بکارگیری PCA و CA در برآورد مدلهای متغیر نهان میباشد.که برای این منظوردو کلاس از مدلهای خطی تعمیم یافته در نظر گرفته میشوند،که در آنها تابع پیش‌بینی کننده یک تابع خطی یا درجه دوم از متغیرهای نهان میباشد.در مدل اول متغیر های پاسخ y_{ij} بوسیله تابع پیش‌بینی‌کننده η_{ij} و بصورت زیر به مدل در آمده اند:

$$\eta_{ij} = \text{Link}(\mu_{ij}) = \alpha_j + \mathrm{x}_i \beta_j \quad ; \quad i=1,\dots,n\,;\, j=1,\dots,m \qquad (۱)$$

بگونه ای که $\mu_{ij} = \mathrm{E}(y_{ij})$ و x_i به عنوان یک عامل ثابت در شکل متغیر نهان در مدل میباشد. توجه ما به توزیعهای نرمال، پواسن و برنولی است. برای توزیع نرمال معادلهٔ (۱) بعضی اوقات یک مدل *آنالیز فاکتور* تابعی نامیده میشود (Anderson-1988)،ودر حالت برنولی نیز با یک مدل دو پارامتری *راش* که در تئوری پاسخ و در روانشناسی کاربرد دارد منطبق میگردد (Andersen -1980). اگر در عمل خطی بودن مورد انتظار نباشد میتوان η_{ij}رایک تابع درجه دو بصورت زیر در نظر گرفت:

$$\eta_{ij} = a_j - \frac{1}{2}\frac{(\mathrm{x}_i - u_j)^2}{t_j^2} \qquad (۲).$$

این مدل و این پارامترگذاری در کاربردهای محیط زیستی متداول است وقتیکه فراوانیهای m گونهٔ مختلف از گیاهان و یا حیوانات (بعوان m ویژگی ماتریس مشاهدات) در n مکان مختلف (بعنوان n نمونه) بصورت تک مدی (Unimodall) بر حسب بعضی عوامل محیطی (متغیر نهان) تغییر میکند.

استفاده از مدلهای متغیر نهان، همچنین در زمینه‌های مطالعاتی اقتصاد، علوم اجتماعی و روانشناسی کاربرد دارد و اغلب، چنین مدلهایی در بهتر دانستن عوامل مؤثر در تغییرات داده‌ها به محقق کمک مینماید. فصل سوم این پایان‌نامه شامل معرفی مدلهای مذکور میباشد.

نویسندگان متعددی به مقایسهٔ CA و PCA با برآورد گر درستنمایی (MLE) و کمترین مربعات در مدلهای متغیر نهان پرداخته‌اند (Gauch, chase, whittaker- 1974)، اما

در مورد ویژگی این برآوردگرها کار عمده‌ای صورت نگرفته است. همچنین مشخص شده است که PCA برای یک مدل خطی نرمال معادل است با MLE و پیشنهادهایی در بکارگیری CA بعنوان تقریب MLE در مدلهای خطی تعمیم یافته نیز ارائه شده است (Ter Braak- 1985). فصل چهارم پایان‌نامه به این موضوع اختصاص دارد. در این فصل به طور مشخص شاخصهای ستونی CA و PCA (که در فصل بعد تعریف خواهند شد) بعنوان برآوردگر پارامترهای مدلهای معرفی شده در نظر گرفته میشوند. زیرا همانطور که خواهیم دید شاخصهای سطری،اطلاعات مربوط به متغیر نهان را در خود دارند، که درتعریف تابعی پیش‌بینی کننده مدل نیازی به دانستن آنها نیست. در فصل پنجم به ارزیابی سازگاری برآوردگرهای حاصل خواهیم پرداخت. یکی از دلایل توجه به سازگاری وجود " مسئله پارامترهای انتشاری" است که در مقدمهٔ این فصل چگونگی آن آمده است.در این فصل سازگار بودن یا نبودن برآوردگرها برای یک تغییر مقیاس ومبدأ از پارامتر ها مورد نظر است.یعنی پاسخ به این پرسش که آیا برآورد ارائه شده در یک ارتباط خطی مجانبی از پارامترها قرار دارد و یا خیر؟.که نتایج حاصل از این بررسی در دو قضیه و یک روش نموداری خلاصه شده اند و بیانگر سازگاری برآورد PCA در حالت نرمال و بشرط وجود گشتاورهای متغیر نهان و ناسازگاری برآوردها در حالتهای دیگر هستند.فصل ششم نیز با شبیه سازی مشاهداتی از مدل(۲)،به مقایسه سازگاری برآوردهای CA و PCA در نمونه ای با حجم معین میپردازدکه در کاربرد بیشتر مورد توجه است.و همچنین در این فصل تعریفی از آمار محیط زیستی و کاربردبخشی از مطالب فوق در این زمینه تحقیقی آمده است.

مباحث این پایان نامه:

فصل دوم:این فصل شامل تعریف و کاربرد تحلیل مؤلفه های اصلی و تحلیل تناظر و مفاهیمی است که میتواند در ارتباط با یک مدل آماری باشد.

فصل سوم:هدف ایـن فصـل معرفـی دو مـدل مشـخص از مدلـهای متغـیر نـهان اسـت کـه بـا فرمولبندی مدلهای خطی تعمعم یافته ارائه میشوندو برآورد پارامتر آنـها در ارتبـاط بـا مطـالب فصل دوم مورد نظر میباشد.

فصل چهارم:در این فصل با تشکیل تابع درستنمایی حاصل از مدلهای معرفی شده بـه بررسـی ارتباط MLE با PCA وCA پرداخته میشود.

فصل پنجم:این فصل با ارائه شاخصهای ستونی PCA وCA بعنـوان بـرآورد پـارامتر مدلـهای مورد نظر به ارزیابی سازگاری آنها نسبت به یک تغییر مقیاس و مبـدأ از پـارامتر هـا و مقایسـه سازگاری در حالتهای گوناگون میپردازد.

فصل ششم:از آنجا که انتخاب مدلهای متغیر نـهان و اسـتفاده از CA در آنـها بـه کـاربردی در تحقیقات محیط زیستی باز میگردد،در این فصل تعریـف داده هـای محیـط زیسـتی و کـاربرد بخشی از موضوعات مطرح شده در فصلهای قبل در ایـن زمینـه تحقیقـی مـورد بررسـی قـرار میگیرد.

فصل دوم

تحلیل مؤلفه‌های اصلی و تحلیل تناظر