بسم الله الرحمن الرحيم

٩٥٩٢٦

دانشگاه پیام نور

مرکز تهران

دانشکده علوم پایه   گروه : آمار

پایان نامه :

برای دریافت درجهٔ کارشناسی ارشد

گرایش: آمار ریاضی

عنوان:

تحلیل رگرسیون معکوس قطعه ای استوار

استاد راهنما:

دکتر مسعود یارمحمدی

۱۳۸۷ /۴/ ۱۳

استاد مشاور:

دکتر مجتبی گنجعلی

۱۳۸۷ /۴/ ۰۳

نگارش :

هادی عبدالحسینی

خرداد ۱۳۸۷

۹۵۹۳۶

«قدر دانی»

با سپاس ایزد دانا و توانا در اینجا بر خود لازم می بینم که از زحمات و همکاری و همفکری اساتید محترم و گرانمایه که طی مراحل مختلف این پایان نامه مرا یاری نموده اند کمال تشکر و قدر دانی خود را ابراز نمایم.

استاد محترم آقای دکتر مسعود یارمحمدی که راهنمایی این پایان نامه را بر عهده داشته و از هیچ کمکی در این مورد دریغ نفرمودند از زحمات و همکاری ایشان نهایت تشکر را می نمایم.

از استاد محترم آقای دکتر مجتبی گنجعلی که در سمت استاد مشاور قبول زحمت فرموده و در انجام این پایان نامه اینجانب را یاری فرمودند نهایت تشکر و سپاسگزاری را دارم.

از اساتید محترم آقای دکتر پرویز نصیری و آقای دکتر علی اکبر رحیم زاده ثانی که داوری این پایان نامه را به عهده گرفته اند کمال تشکر را دارم.

در پایان هم از تمامی دوستانی که مرا در تهیه منابع و مراجع این پایان نامه یاری رسانده اند بویژه آقای رضا ندیمی کمال تشکر و امتنان را دارم.

تقدیم:

به آنان که پیمانهٔ «لا» زدند .....

فهرست مندرجات

فصل اول:کلیات و مسائل مربوط

فصل دوم: رگرسون معکوس قطعه ای

فصل سوم: روشهای استوار با نقطهٔ فروریزش بالا

فصل چهارم: رگرسیون معکوس قطعه ای استوار

چکیده

روش رگرسیون معکوس قطعه ای، روشی است که برای کاهش بعد، در مدلهای رگرسیونی استفاده می شود. عمل کاهش بعد براساس اطلا عاتی که در منحنی رگرسیون معکوس قرار می گیرد انجام می شود. در مراحل مختلف روش رگرسیون معکوس قطعه ای از برآوردگرهای کلاسیک استفاده می شود. بنابراین نتایج روش در مقابل داده های دور افتاده نااستوار است. یک روش برای استوار کردن این روش جایگزینی برآوردگرهای کلاسیک با برآوردگرهای استوار است. در این رساله روش رگرسیون معکوس قطعه ای به عنوان یک روش کاهش بعد معرفی شده و سپس با ارائهٔ یک مثال کاربردی نحوهٔ استفاده از این روش مورد بررسی قرار می گیرد. در پایان به استوار سازی این روش با استفاده از برآودگرهای استوار می پردازیم.

واژه های کلیدی: کاهش بعد، رگرسیون معکوس قطعه ای، دور افتاده ها، نقطهٔ فروریزش، استواری

فصل ۱

کلیات و مسائل مربوط

۱– ۱ مقدمه

با توجه به اینکه یک مسئلهٔ ویژه در تحلیل داده ها، فزونی بیش از حد اطلاعات[1] است و این فزونی بیش از حد اطلاعات از تعداد زیاد متغیرهای مورد استفاده برای مشاهدات در مجموعهٔ داده ها می آید، و با توجه به اصل امساک که ما را از استفادهٔ بیش از حد متغیرها منع می کند، روشهای گوناگونی در رابطه با کاهش تعداد متغیرها و به نوعی در کاهش بعد آمده است که به ما کمک می کند با استفاده از کمترین تعداد متغیر به حداکثر اطلاعات دسترسی پیدا کنیم. این مسئلهٔ فزونی بیش از حد اطلاعات اغلب در مدلهای وابسته – مستقل نمایان می شود. یعنی زمانی که فرد می خواهد وابستگی متغیر پاسخ Y را با یک یا بیشتر متغیر پیشگوی X_1، X_2، ...، X_m بدست آورد.

[1] Overabundance of Information

برای $m > ۲$ این مسئله اغلب بدون استفاده از برخی صورتهای کاهش بعد غیرممکن می شود. در تحلیل داده ها، این مسئله اغلب بوسیلهٔ ساختن تصاویر پایین بعدی از داده ها حل می شود. این تصاویر ابعاد پایین تر، درمدلهایی استفاده می شوند که متغیر پاسخ، تابعی ازاین تصاویر پایین بعدی از فضای متغیرهای کمکی است. این امیدواری وجود دارد که رابطهٔ بین پاسخ و متغیرهای کمکی در فضای پایین بعدی حفظ شود. معمولاً تابعی که آماردانها انتخاب می کنند و یا ترجیح می دهند انتخاب کنند، تابعهای خطی است که آن هم به علت سادگی و تفسیر پذیری آنهاست. با این حال برخی مواقع یک ساختار صحیح از رابطهٔ بین پاسخ و متغیرهای کمکی نمی تواند به طورمناسبی بوسیلهٔ یک تابع خطی حفظ شود. روشهای متفاوت بسیاری تهیه شده اند که این امیدواری را می دهند که اجازهٔ انعطاف بیشتری را برای به مدل درآوردن این رابطه، با حفظ سادگی و تفسیر پذیری ایجاب کنند، که از جملهٔ این روشها می توان به رگرسیون پیگیری تصویر[۲] اشاره کرد. از روشهای اخیر هم می توان به روش رگرسیون معکوس قطعه ای(SIR)[۳] اشاره کرد.

۱-۲ تاریخچه ، ضرورت و کاربردهای روش

روش رگرسیون معکوس قطعه ای و کاربردهای آن نخستین بارتوسط لی[٤] در سال ۱۹۹۱ مطرح شده است. روش رگرسیون معکوس قطعه ای روشی سودمند برای کاهش بعد در مسئلهٔ رگرسیون ناپارامتری است. این روش هنگامی استفاده می شود که متغیر پاسخ Y بستگی به k ترکیب خطی نامشخص از متغیرهای توضیحی $X = (X_1, ..., X_m)$ داشته و صورت دقیق و کامل وابستگی معلوم نباشد. لازم به ذکراست که روش رگرسیون معکوس قطعه ای برای کاهش بعد متغیرهای توضیحی، بدون انجام یک فرایند برازش مدل پارامتری یا ناپارامتری به کار می رود.

---
[۲] Projection Pursuit Regression
[۳] Sliced Inverse Regression
[٤] Li

در مورد کاربردهای این روش می توان به مباحث کاربردی آمار مانند پزشکی ، ژنتیک و زیست شناسی که درآنها عموماً با داده های با ابعاد بالا که روشهای معمولی رگرسیونی معمولاً قادر به انجام آن نبوده و تحلیل داده ها در ابعاد بالا آسان نیست، اشاره کرد. با استفاده از روش رگرسیون معکوس قطعه ای می توانیم ضمن کاهش بعد داده ها، به طور چشم گیری در تحلیل این داده ها به نتایج دقیق تری دست یابیم.

۱–۳ داده های دورافتاده در مسائل رگرسیونی

یک تعریف کلی برای مشاهدات دورافتاده به اینصورت بیان می گردد که مشاهدهٔ دورافتاده مشاهده ای است که ماندهٔ آن از نظر قدرمطلق خیلی بزرگتر از مانده های سایر مشاهدات است. البته گاهی مشاهدات دورافتاده به گونه ای دیگر خود را بروز می دهند در برخورد با داده ها برای شناسایی نقاط دورافتاده، معیارهایی وجود دارند که فرد با استفاده ازآنها می تواند به دسته بندی داده ها اقدام کند. بدین معنی که برخی نقاطی که از اکثریت داده ها متفاوت هستند و به نوعی می توان آنها را ((خاص)) نامید شناسایی کرد. یکی از معروف ترین این معیارها مربوط به نقاط با نفوذ می باشد. این نقاط مهم هستند چرا که در برازش مدلها و در نتیجه در مقادیر پیش بینی شده تأثیر می گذارند. نکته ا ی که در اینجا بایستی به آن توجه داشت این است که لزوماً همهٔ نقاط با نفوذ بالا دادهٔ دورافتاده نیستند و بدین مفهوم نیست که آنها را باید از داده ها جدا کرده و یا حذف نمود. چه بسا که بعضی از این نقاط حاوی اطلاعات مفیدی در مورد جامعهٔ مورد بررسی باشند. در رگرسیون خطی دور افتادگی در چندین روش تعریف می شود.

برای مدل رگرسیونی در حالت چند متغیره که به صورت زیر بیان می شود:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon \qquad (1.3.1)$$

که X_1، X_2، ...، X_m متغیرهای توضیحی و β_0، β_1، ...، β_m ضرایب رگرسیونی در حالت چند متغیره هستند. نفوذ i امین نقطه در فضای چند متغیّره به صورت زیر تعریف می شود:

$$h_{ii} = x_i'(X'X)^{-1} x_i \qquad (۲.۳.۱)$$

که X ماتریس متغیّر کمکی $n \times k$ است. نفوذ یک نقطه به طور ویژه بیان می کند که چقدر یک نقطهٔ خاص x_i از مرکز تودهٔ متغیّرهای مستقل، یعنی از $\bar{x} = \frac{1}{n} I'X$ دور است(که I برداری n بعدی با مؤلفه های واحد است). در مورد توده های تک مُدی از نقاط متغیّر های کمکی این می تواند یک اندازهٔ مناسب دوری یک نقطه ازمرکز اکثریت نقاط باشد.

مسئله ای که در اینجا مورد توجه است این است که داده هایی که در نهایت داده دورافتاده هستند می توانند اثرات بدی روی برآوردگرها و خواص آنها داشته باشند. بنابراین معرفی برآوردگرهایی که بتوانند در مقابل داده های دورافتاده مقاومت نشان داده و کمتر تحت تأثیر این داده ها واقع شوند می تواند بسیار حائز اهمیت باشد. این برآوردگرها در قالب برآوردگرهای استوار[5] معرفی می شوند.

بنابراین پس از ارائهٔ یک برآوردگر، مسئله اساسی رفتار آن برآوردگر در رابطهٔ با داده های دورافتاده و بررسی اثرات مخرب آنهاست. در مرحلهٔ بعد به منظور مقاوم سازی برآوردگر در مقابل داده های دورافتاده استفاده از روشهای استوار سازی لازم می باشد.

حال با توجه به اینکه جهت بدست آوردن برآوردگر *SIR* از معیارهای میانگین و تابع اتو کوواریانس استفاده می شود که به داده های دور افتاده مقاوم نمی باشند، لذا با استفاده از روشهای استوار سازی، به اصلاح این برآوردگرها پرداخته و سعی می کنیم آنها را در برابر داده های دور افتاده مقاوم تر نمائیم.

در ادامهٔ این فصل با ارائهٔ برخی نمادها و تعریف ها یی که در فصول آینده به آنها نیاز داریم، که از جملهٔ آنها تعاریف و مباحث مربوط به استواری برآوردگرها را می توان نام برد، می پردازیم.

---

[5] Robust estimators

۱-٤ هم وردایی[٦]

مفهوم هم وردایی یک برآوردگر بصورت زیر تعریف می شود:

برآوردگر T هم ورداست اگر هم وردای رگرسیونی،[۷] هم وردای مقیاسی[۸] و هم وردای وابسته[۹] باشد.

برآوردگر Tرا هم وردای رگرسیونی گوییم هر گاه :

$$T(\{(x_i, y_i + x_i v) ; i = 1,\ldots,n\}) = T(\{(x_i, y_i) ; i = 1,\ldots, n\}) + v \qquad (1.4.3)$$

برآوردگر T را هم وردای مقیاسی گوییم هرگاه :

$$T(\{(c\, x_i, y_i + x_i v) ; i = 1,\ldots,n\}) = c\, T(\{(x_i, y_i) ; i = 1,\ldots, n\}) \qquad (1.4.4)$$

برآوردگر T را هم وردای وابسته گوییم هر گاه :

$$T(\{(x_i \mathbf{A}, y_i + x_i v) ; i = 1,\ldots,n\}) = \mathbf{A}^{-1}\, T(\{(x_i, y_i) ; i = 1,\ldots, n\}) \qquad (1.4.5)$$

که در آن $\mathbf{A}$ هر ماتریس ناویژهٔ مربع است.

باید توجه داشت برای این که یک برآوردگر هم وردا باشد باید در هر سه تعریف هم وردایی مطرح شده در فوق صدق کند و صرف هم وردای رگرسیونی یا هم وردای مقیاسی یا هم وردای وابسته بودن و یا دو مورد از موارد فوق را دار بودن نمی تواند به هم وردا بودن برآوردگر منجر شود.

۱-٥ نقطهٔ فروریزش[۱۰] :

فرض کنید یک نمونه شامل n مشاهدهٔ :

$$z = \{(x_{11}, \ldots, x_{1m}, y_1), \ldots, (x_{n1}, \ldots, x_{nm}, y_n)\}$$

[٦] Equivariant
[۷] Regression equivariant
[۸] Scale equivariant
[۹] Affine equivariant
[۱۰] Break down point

و T برآوردگر رگرسیونی برای پارامتر θ باشد دراینصورت با به کار بردن T و نمونهٔ z ، حاصل یک بردار از ضرایب رگرسیونی $\hat{\underline{\theta}}=\begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\\ \vdots\\ \hat{\theta}_m\end{pmatrix}$ می باشد به طوری که $T(z)=\hat{\underline{\theta}}$. حال اگر تمام نمونه های مخدوش شدهٔ z' که بوسیلهٔ جایگذاری m مقدار دلخواه به جای مقادیر اصلی بدست آمده را در نظر بگیریم، بیشترین اریبی که می تواند توسط ورود یک ناخالصی ایجاد شود عبارت است از :

$$bias\ (m,T,z) = sup_{z'} \left\| T(z') - T(z) \right\| \qquad (٦.٥.١)$$

اگر $bias(m,T,z)$ نامتناهی باشد، بدین معنی است که m نقطهٔ دورافتاده روی برآوردگر T اثر زیاد داشته و در نتیجه برآوردهای حاصل نامناسب خواهند بود.

برای یک نمونهٔ متناهی z، نقطهٔ فروریزش برآوردگر T توسط رابطهٔ زیر تعریف می شود :

$$\varepsilon_v^* = \left\{ min \frac{m}{n} \quad s.t \quad bias(m;T,z) = \quad \infty \right\} \qquad (٧.٥.١)$$

یعنی $\min \frac{m}{n}$ به طوری که $bias(m;T,z)$ نامتناهی باشد.

به عبارت دیگر این کوچکترین کسری از ناخالصی داده هاست که سبب می شود برآورد T مقداری دورتر از مقدار مورد انتظار اختیار کند. در روش کمترین توان های دوم مشاهده می شود که تنها یک نقطهٔ دورافتاده کافی است که این برآوردگر مغشوش شود. بنابراین نقطهٔ فروریزش برآوردگر روش توان های دوم $\frac{1}{n}$ است. که اگر حجم نمونه افزایش یابد عبارت فوق به سمت صفر میل می کند. بنابراین می توان گفت که روش توان های دوم دارای نقطهٔ فروریزش صفر درصد می باشد.

تعریف دیگری از نقطهٔ فروریزش به صورت زیر بیان می گردد(کسلّا و برگر[۱۱] ۱۹۹۰، صفحهٔ ۲٤۲):

اگر $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ یک نمونهٔ مرتب شده با حجم n باشد و T_n یک آماره براساس این نمونه باشد، آنگاه T_n دارای نقطهٔ فروریزش b است که $0 \leq b \leq 1$، هر گاه داشته باشیم:

$$\forall \varepsilon > 0 \quad \lim_{x_{((1-b)n)} \to \infty} T_n < \infty \qquad \lim_{x_{(1-(b+\varepsilon)n)} \to \infty} T_n = \infty \qquad (۸.۵.۱)$$

با توجه به روابط بالا می توان نشان داد که نقطهٔ فروریزش میانگین صفر درصد و نقطهٔ فروریزش میانه پنجاه درصد است.

نقاط فروریزش نقش مهمّی در برآوردهای استوار دارند. با توجه به تعریف نقطهٔ فروریزش هر چه یک برآورد گر دارای نقطهٔ فروریزش بالایی باشد آن برآوردگر در برابر نقاط دورافتاده کارایی بالایی دارد. بنا بر این جهت کاهش اثر نقاط دورافتاده بر روی برآورد پارامتر ها، یافتن برآوردهای با نقاط فروریزش بالا ضروری می باشد.

۱-٦ نقطهٔ فروریزش نمونه متناهی[۱۲]

این نماد نخستین بار در سال ۱۹٦۷ و بوسیلهٔ هاج[۱۳] منتشر شد، و سپس بوسیلهٔ همپل[۱٤] (۱۹٦۸،۱۹۷۱) تعمیم پیدا کرد. نسخهٔ نمونه متناهی از نقطهٔ فروریزش شامل نقطهٔ فروریزش جمعی[۱٥] (ABP) و نقطهٔ فروریزش جایگزین[۱٦] (RBP) است. که بوسیلهٔ دوناهو و هوبر[۱۷] (۱۹۸۳) معرفی شده و برا ی ارزیابی استواری برآوردگرها استفاده شده که در زیر چگونگی بدست آوردن آنها نشان داده می شود.

---

[۱۱] Casella & Berger
[۱۲] Finite sample breakdown point
[۱۳] Hodge
[۱٤] Hample
[۱٥] Addition breakdown point
[۱٦] Replacement breakdown point
[۱۷] Donoho and Huber

فرض کنید $X^n = \{X_1, \ldots, X_n\}$ یک نمونهٔ بـه انـدازهٔ n در R^d باشـد. نقطـهٔ فروریـزش جمعی نمونه متناهی از یک برآوردگر T در X^n به صورت زیرتعریف می شود :

$$ABP(T, X^n) = \min\left\{\frac{m}{m+n} : \sup_{Y^m} \left\|T(X^n \cup Y^m) - T(X^n)\right\| = \infty\right\} \quad (۹.٦.۱)$$

که Y^m دلالت بر مجموعه داده هایی از اندازهٔ n با مقادیر دلخواه و $X^n \cup Y^m$ دلالت بر نمونـهٔ آلوده شده بوسیلهٔ مقادیر Y^m اضافه شده به X^n می باشد.

نقطهٔ فروریزش جایگزینِ نمونه متناهی از یک برآوردگر T در X^n به صورت زیر تعریـف می شود :

$$RBP(T, X^n) = \min\left\{\frac{m}{n} : \sup_{X^n_m} \left\|T(X^n_m \cup Y^m) - T(X^n)\right\| = \infty\right\} \quad (۱۰.٦.۱)$$

که X^n_m دلالت بر نمونهٔ آلوده شده از X^n را داردکه بوسیلهٔ جایگزینی m نقطه از X^n با مقادیر دلخواه، ایجاد شده است.

به بیان دیگر ABP و RBP یِ یک برآوردگر به ترتیب کمترین کسرِ اضـافه شـده و کـسرِ جایگزینی است که می تواند اریبی برآوردگر را به سمت بینهایت ببرد.

۱–۷ روابط کمی بین ABP و RBP

با توجه به ABP و RBP یِ معرفی شده در بخش ۱–٦، ایـن دو تعریـف از نقطـهٔ فروریـزش اغلب به طور جداگانه برای برآوردگرها مورد بحث واقع مـی شـوند و بـه صـورت دو مفهـوم کاملاً مستقل رفتار می کنند. برخی نویسنده ها علاقه مند بـه اسـتفاده از ABP در بحـث هـای استواری برآوردگرها هستند در حالی که برخی دیگر RBP را برای بررسی خـواص اسـتواری ترجیح می دهند و اعتقاد دارند که ساده تر، واقع گرایانه تر و به طور کلی کاربردی تـر اسـت. برای این که بتوان به قضاوت کامل تری در مورد استواری برآوردگرهـا دسـت یافـت بـا بیـان روابط کمی بین این دو در کلاس بزرگی از برآوردگرها می توان با بدست آوردن اندازهٔ یکی از دو معیار دیگری را نیز بدست آورد.

یک نکتهٔ جالب در مورد نقاط فروریزش یک برآوردگر T در X^n این است که در بسیاری از موارد، آنها فقط به اندازهٔ نمونهٔ n وابسته هستند و از ترکیب و پیکربندی X^n مستقل هستند (هر چند از تعاریف این چنین به نظر می آید که آنها بایستی به نمونهٔ X^n وابسته باشند). به بیان دیگر برای نقاط فروریزش که در حالت کلی در ارزیابی استواری یک برآوردگر استفاده می شوند این خاصیتِ «آزاد از نمونه» به طور قطع مطلوب است.

در زیر توجه خود را به نقاط فروریزشی منحصر می کنیم که از شکل X^n مستقل هستند و RBP هایی که به فرمِ $\frac{\lfloor an+b \rfloor}{n}$ یا $\frac{\lceil an+b \rceil}{n}$ هستند. در واقع این مورد برای بیشتر برآوردگرهای مکانی و پراکنش صادق است. دراینجا $\lfloor x \rfloor$، نشان دهندهٔ بزرگترین عدد صحیح کوچکتر یامساوی x و $\lceil x \rceil$ کوچکترین عدد صحیح بزرگتر یا مساوی x است.

قضیه ۱-۱ :

فرض کنید T یک برآوردگر در R^d باشد و X^n یک مجموعه داده با اندازهٔ n باشد.

الف) نقطهٔ فروریزش جمعی T می تواند به صورت زیر بیان شود :

$$ABP(T,X^n) = (\lfloor \varepsilon n + \beta \rfloor + m)/(n + \lfloor \varepsilon n + \beta \rfloor + m)$$

اگر نقطهٔ فروریزش جایگزینِ T بتواند بصورت زیر بیان شود:

$$RBP(T,X^n) = \lfloor (\varepsilon n + (\beta + m + \varepsilon))/(\varepsilon + ۱) \rfloor / n$$

ب) نقطهٔ فروریزش جایگزین T می تواند به صورت زیر بیان شود :

$$RBP(T,X^n) = \lfloor (\varepsilon n + (\beta + m + r))/(\varepsilon + ۱) \rfloor / n$$

اگر نقطهٔ فروریزش جمعی T بتواند بصورت زیر بیان شود:

$$ABP(T,X^n) = (\lfloor \varepsilon n + \beta \rfloor + m)/(n + \lfloor \varepsilon n + \beta \rfloor + m)$$

که ε و β و عدد صحیح m برخی ثابتهایی هستند که از n و X^n مستقل هستند $۰ \le \beta \le \varepsilon \le ۱$، و

$$\max\{۱ - (\varepsilon n + \beta + m)/(\varepsilon + ۱), \max_N \{D^-(N)\}\} \le r/(\varepsilon + ۱) < ۱ + \min_N \{D^-(N)\}$$

با

$N = \lfloor (\varepsilon + ۱)n + \beta + m \rfloor$ و

$$D^{-}(n)=\lfloor(\varepsilon n+\beta+m)/(\varepsilon+۱)\rfloor-(\varepsilon n+\beta+m)/(\varepsilon+۱)$$

اثبات : زو[۱۸] ۱۹۹۹ راببینید.

قضیه ۱-۲ :

فرض کنید T یک برآوردگر در R^d باشد و X^n یک مجموعه با اندازهٔ n باشد .

الف) نقطهٔ فروریزش جمعی T می تواند به صورت زیر بیان شود :

$$ABP(T,X^n)=(\lceil\varepsilon n+\beta\rceil+m)/(n+\lceil\varepsilon n+\beta\rceil+m)$$

اگر نقطهٔ فروریزش جایگزین T بتواند به صورت زیر نوشته شود:

$$RBP(T,X^n)=\lceil(\varepsilon n+\beta+m+\varepsilon)/(\varepsilon+۱)\rceil/n$$

ب) نقطهٔ فروریزش جایگزین T می تواند به صورت زیر نوشته شود

$$RBP(T,X^n)=\lceil(\varepsilon n+(\beta+m+r))/(\varepsilon+۱)\rceil/n$$

اگر نقطهٔ فروریزش جمعی T بتواند به صورت زیر نوشته شود

$$ABP(T,X^n)=(\lceil\varepsilon n+\beta\rceil+m)/(n+\lceil\varepsilon n+\beta\rceil+m)$$

که ε و β و عدد صحیح m برخی ثابتهایی هستند که از n و X^n مستقل هستند $۰\le\beta\le\varepsilon\le۱$، و

$$\{\max\{-(\varepsilon n+\beta+m)/(\varepsilon+۱),\max_N\{D^{+}(N)\}-۱\}\le r/(\varepsilon+۱)<\min_N\{D^{+}(N)\}$$

با $N=\lfloor(\varepsilon+۱)n+\beta+m\rfloor$ و

$$D^{+}(n)=\lceil(\varepsilon n+\beta+m)/(\varepsilon+۱)\rceil-(\varepsilon n+\beta+m)/(\varepsilon+۱)$$

اثبات : زو ۱۹۹۹ راببینید.

مثال ۱-۱:

RBP ی میانهٔ نمونه ای را در $R^۱$ با استفادهٔ مستقیم از ABP ی آن بدست می آوریم.

ABP ی میانهٔ نمونه ای در دوناهو و هوبر (۱۹۸۳) داده شده است که آن مقداری برابر با $\frac{۱}{۲}$ است. برای استفاده نمودن از قضیهٔ ۱-۱ نیاز است که ε و β و m را مشخص نمائیم. آنها به صورت زیر در نظرگرفته می شوند $\beta=۰$ و $\varepsilon=۱$.

[۱۸] Zou

حال با استفاده از قضیه۱-۱، RBP یِ برآوردگر میانهٔ نمونه ای برابر $\lfloor (n+r)/2 \rfloor / n$ می باشد به طوری که :

$$\max\left\{\frac{1}{2}, \max_N \{\lfloor N/2 \rfloor - N/2\}\right\} \leq r/2 < 1 + \min_N \{\lfloor N/2 \rfloor - N/2\}$$

و $N = \lfloor 2n \rfloor = 2n$. بنابراین نتیجه می گیریم که $1 \leq r < 2$. لذا، RBP یِ برآوردگر می تواند به صورت $\lfloor (n+1)/2 \rfloor / n$ نوشته شود، که می توان نشان داد این RBP در واقع همانند RBP یی است که از تعریف اصلی RBP مشتق می شود.

مثال ۱-۲ :

ABP یِ L_1- میانه که در نوشته ها میانهٔ فضایی هم نامیده می شود را در R^d، $(d \geq 1)$ به طور مستقیم از RBP بدست می آوریم. L_1- میانه به صورت بردار T تعریف می شود به طوریکه :

$$T = \arg\min_{x \in R^d} \sum_{i=1}^{n} \| X_i - x \| \qquad (11.7.1)$$

توجه شود که منظور از $argmin$ یِ مجموع فوق، $x \in R^d$ یی است که به ازای آن مجموع کلیهٔ فواصل اقلیدسی X_i از x به حداقل برسد.

برای هر نمونهٔ $X_1, \ldots, X_n$ لوپا و روسو[19] (۱۹۹۱) نشان داده اند که :

$$RBP(T, X^n) = \lfloor (n+1)/2 \rfloor / n$$

برای به کار بردن قضیهٔ ۱، نیاز داریم که ε ، β و m را مشخص کنیم. به نظر می رسد برای این برآوردگر $m = 0$ ، $\beta = 0$ و $\varepsilon = 1$ به طور منحصر به فردی مشخص می شوند. حال با استفاده از قضیهٔ ۱ داریم که ABP یِ L_1- میانه $\frac{1}{2}$ است.

۱-۸ توزیع بیضوی

$F_{\mu,\Sigma}$ دارای توزیع بیضوی است هر گاه چگالی آن به فرم زیر باشد:

$$f(\mathbf{x}) = (\det(\Sigma))^{\frac{-1}{2}} f\left[(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right] \qquad (12.8.1)$$

[19] Lopuha? and Rousseeuw

که $f:[\cdot,\infty]\to[\cdot,\infty]$ یک تابع معین است و $(\mu,\Sigma)\in\Theta$ که $\Theta=R^p\times PDS(p)$ و PDS[۲۰] ماتریس متقارن معین مثبت است.

۱–۹ فضای پوچ

فضای پوچ یک ماتریس $m\times n$ مانند A، به صورت مجموعهٔ زیر تعریف می شود:

$$Null(\mathbf{A})=\{X\in R^n:\mathbf{A}X=o\}$$

که o دلالت بر بردار صفر با m مؤلفه دارد. معادلهٔ ماتریسی $\mathbf{A}X=o$ معادل ، معادلات همگن زیر می باشد:

$$\mathbf{AX}=o\Leftrightarrow\begin{array}{c}a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n=\cdot\\a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n=\cdot\\\vdots\\a_{m1}x_1+a_{m2}x_2+\cdots+a_{mn}x_n=\cdot\end{array}$$

با این نمایش فضای پوچ حل مجموعهٔ معادلات همگن فوق می باشد.

۱–۱۰ تحلیل همبستگی کانونی

تحلیل همبستگی کانونی به شناخت و کمی کردن رابطهٔ بین دو مجموعه از متغیرها پرداختـه، اولین بار توسط هتلینـگ (۱۹۳۵) مطـرح شـد. در اصـل تحلیـل همبـستگی کـانونی در مـورد همبستگی بین یک ترکیب خطی از متغیرهای یک مجموعه و یک ترکیـب خطـی از متغیرهـای مجموعهٔ دیگر متمرکز می شود. ابتدا هدف این است که دو ترکیب خطی با بیشترین همبستگی تعیین شود. سپس دو ترکیب خطی را تعیین می کنیم که در میان تمـام زوجهـای ناهمبـسته بـا زوج انتخاب شدهٔ اول دارای بیشترین همبستگی باشد. زوجهای ترکیبـات خطـی را متغیرهـای کانونی و همبستگی آنها را همبستگی های کانونی می نامنـد. همبـستگی هـای کـانونی شـدت ارتباط بین دو مجموعه ازمتغیرها را اندازه می گیرد. آنچه تحلیل همبستگی کـانونی انجـام مـی دهد این است که روابط بین دو مجموعه را که براساس تعداد زیاد متغیرهـا پایـه ریـزی شـده

[۲۰] Positive Definite Symmetric