دانشگاه پیام نور تهران
دانشکده علوم
گروه آمار

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی آمار گرایش آمارریاضی

عنوان:
کاربرد رگرسیون مؤلفه‌های اصلی برای مجموعه داده‌های با مشاهدات دورافتاده و گم شده

استاد راهنما:
دکتر مسعود یارمحمدی

استاد مشاور:
دکتر علی شادرخ

پژوهشگر:
راضیه فرمانی اردهائی

خرداد ۹۰

چکیده

در این تحقیق به بررسی داده های با بعد بالا، که شامل داده های دورافتاده و گمشده می باشند، پرداخته می شود. روش های مختلف رگرسیون مولفه های اصلی که می‌توانند در صورت وجود داده‌های دور افتاده و گم شده استوار باشند بررسی می شود. سپس رگرسیون مولفه های اصلی استوار به عنوان روش رگرسیونی که بخاطر بعد بالا در متغیرهای پاسخ و پیشگو از روش تعقیب تصویر و تحلیل مولفه های اصلی برای کاهش بعد استفاده می کند ، و در برابر داده های دور افتاده وگم شده استوار می باشند را شرح می‌دهیم. سپس با استفاده از روشهای شبیه سازی برازش مدلهای رگرسیونی به داده هایی که با داده های دور افتاده و گم شده آلوده شده اندرا مورد بحث قرار می دهیم. در پایان با مقایسه روش های مختلف رگرسیونی برازش شده به داده های شبیه سازی شده، روش استوارتر در مقابل داده های دور افتاده و گم شده معرفی می‌گردد.

کلید واژه: مولفه های اصلی- داده‌های دور افتاده و گم شده- رگرسیون مولفه های اصلی- تحلیل مولفه های اصلی- روش های استوار

پیش گفتار

مدل‌های رگرسیونی به بررسی رابطه بین یک یا چند متغیر پیشگو و متغیر پاسخ می‌پردازند. در مدل رگرسیونی خطی چندگانه، روش کمترین توان‌های دوم LS به علت سادگی در امر محاسبات یکی از کاربردی‌ترین روش‌های برآورد پارامترهای مدل رگرسیونی محسوب می‌شود. برآوردهای روش کمترین توان‌های دوم در صورت وجود نقاط دورافتاده در متغیرهای پاسخ و یا متغیرهای توضیحی منجر به نتایج نادرست و غیر دقیق شده لذا استفاده از روشهای استوار جهت خنثی نمودن و کاهش اثر این نقاط پیشنهاد می شود.

در رگرسیون یک متغیره بر اساس ایده همپل روسو(۱۹۸۴) برآوردگر کمترین توان‌های دوم خطا و همچنین روسو(۱۹۸۴) برآوردگر کمترین توان‌های دوم پیراسته[۱](LTS) و برآوردگر کمترین توان‌های دوم دوباره وزن دار شده[۲](RLS) را به عنوان برآوردگرهای رگرسیونی استوار پیشنهاد کرد. مارونا (۱۹۷۶) برآوردگرهای M را معرفی کرد. در سال ۱۹۸۶ مارونا و مورگن تالر[۳] از برآوردگرهای چندمتغیره M در محاسبه برآوردهای رگرسیونی استفاده کردند.

رگرسیون مؤلفه اصلی[۴] (PCR) نیز یکی از معمول‌ترین روشهای رگرسیونی است که بین دو مجموعه از داده‌ها که یکی شامل متغیرهای پیشگو و دیگری شامل متغیر(های) پاسخ می‌باشد یک رابطه خطی می‌برازد. معمولا روش رگرسیون مولفه های اصلی هنگامی که متغیرهای پاسخ در بعد کم و پیشگوها نسبت به تعداد مشاهدات دارای بعد زیاد هستند مورد استفاده قرار می گیرد. همچنین روش

۱- Least Trimmed Squares
۲- Reweighted Least Squares
۳-Maronna & Morgenthaler
۴ -Principal Component Regression

رگرسیون مولفه اصلی هنگامی که مشکل همبستگی در متغیرهای توضیحی(همخطی) در داده ها کـه یکی از اساسی ترین مشکلات در برآوردهای حداقل مربعـات در رگرسـیون چندگانـه اسـت نیـز قابـل کاربرد است. زیرا اگر همخطی در داده ها موجود باشد آنگاه واریانسهای بعضـی از برآوردهـای ضـرایب رگرسیونی ممکن است خیلی بزرگ شوند و به برآوردهـای غیرقابـل اطمینـانی از پارامترهـای معادلـه رگرسیون منجر شوند برای غلبه بر این مشکل روشهای زیادی پیشنهاد شده یک راه ممکن این اسـت که تنها زیرمجموعه ای از متغیرهای پیشگو را استفاده کنیم به طوریکه این مجموعـه طـوری انتخـاب می‌شود که همخطی وجود نداشته باشد. راه‌های متعددی برای انتخاب زیرمجموعه وجـود دارد کـه در بین آنها روشهایی بر اساس مولفه‌های اصلی می باشد. رگرسیون مولفه‌های اصلی به سادگی با استفاده از مولفه‌های اصلی متغیرهای پیشگو به جای متغیرهای پیشگوی اولیه شـروع مـی شـود. از آنجاییکـه مولفه‌های اصلی ناهمبسته هستند هیچگونه همخطی بین آنها موجود نیست و محاسـبات رگرسـیونی نیز ساده‌تر خواهد بود. اگر همهٔ مولفه‌های اصلی در مدل به کارروند، آنگاه مدل به دست آمـده معـادل است با مدل به دست آمده از روش حداقل مربعات است بنابراین واریانسهای بـزرگ نـا شـی از وجـود همخطی از بین نمی‌رود. اگر برخی از مولفه‌های اصلی از معادله رگرسیونی حذف شوند برآوردگرهـایی برای ضرایب معادله رگرسیون اصلی به دست می آیند. که این برآوردگرها معمولاً اریبند ولی همزمـان می‌توا نند واریانس بزرگ برآوردگرهای ضرایب رگرسیونی را که ناشی از وجود همخطی است، کـاهش دهند.

در فصل اول این تحقیق به تعریف داده های گمشده و داده های دور افتاده و بیان روش های استوار و مفاهیم مرتبط با آن نظیر نقطه فروریزش و خاصیت هم وردایی نسبی می پـردازیم. در فصـل دوم بـه علت این که در این تحقیق متغیر پاسخ و متغیرهای توضیحی با بعد بالا در نظر گرفتـه مـی شـود بـه شرح روش های تحلیل مولفه های اصلی[۵] که روشی معروف برای کاهش بعـد اسـت مـی پـردازیم. در فصل سوم به معرفی روش های رگرسیون مولفه‌های اصلی که می توانند در مقابل داده های دورافتـاده

۱-Principal Component Analysis

و گمشده مقاومت کنند، و به همین علت روش های رگرسیون مولفه‌های اصلی استوار[۶] نامیده می شوند می پردازیم. در فصل چهارم با کمک شبیه سازی روش های رگرسیون مولفه‌های اصلی استوار بحث شده در فصل سوم با روش رگرسیون مولفه‌های اصلی کلاسیک مقایسه می‌شود.

۲-Robust Principal Component Regression

فهرست مندرجات

عنوان **صفحه**

پیوست ها

فصل اول

تعاریف وپیش نیازها

۱-۱ داده های گمشده

داده های گمشده نتیجه بی پاسخی یعنی عدم موفقیت دراندازه گیری بعضی از واحدهای منتخب می باشد. زمانی ایجاد می شود که یک واحد از واحدهای آماری به تمام یا قسمتی از پرسش های پرسشنامه‌ی آمارگیری پاسخ ندهد.دلایل بی پاسخی عبارتند از:

۱- عدم دسترسی به پاسخگو: عدم دسترسی به پاسخگو میتواند ناشی از مراجعه در وقت نامناسب، جابجایی بعضی از واحدهای آماری و ... باشد. می توان بوسیله تنظیم زمان و تعداد دفعات مراجعه، صحیح وکامل بودن آدرس هر یک از واحدهای آماری چنین داده های گمشده ای را کاهش داد.

۲- عدم همکاری پاسخگو: در این مرحله به واحد آماری دسترسی پیدا کرده ایم اما پاسخگو از مشارکت در مصاحبه و یا در صورت شرکت در مصاحبه از پاسخ دادن به سئوالات خاصی خودداری می نماید.روشهای کاهش عدم همکاری پاسخگوعبارت ازتغییر در نحوه برخورد آمارگیر در زمان محاسبه، اطمینان پاسخگو از محرمانه ماندن اطلاعات و...است.

۳ - عدم توانایی و شایستگی پاسخگو

۱-۱-۱ ساختارهای گم شدگی

در این بخش به معرفی ساختارهای گم شدگی می پردازیم.

متغیر نشانگر پاسخ R_{it} به صورت زیر تعریف می کنیم($t = ۱,۲$):

اگر متغیر پاسخ برای فرد i در دوره t مشاهده شود $R_{it} = ۱$ و اگر Y_{it} گمشده باشد $R_{it} = ۰$

فرض کنید که $Y_{i۱}^{\circ}$ نمایانگر مقدار مشاهد شده‌ی واحد iام در دور اول آمارگیری و $Y_{i۲}^{m}$ مقدار گم شدگی واحد iام در دوره دوم آمارگیری باشد. احتمال گم شدگی واحد iام در دور دوم آمارگیری به شرط مقدار پاسخ مشاهده شده‌ی $Y_{i۱}^{\circ}$ و مقدار پاسخ گم شده‌ی $Y_{i۲}^{m}$ به صورت زیر نشان داده می شود:

$$P(R_{i۲} = ۰ \mid Y_{i۱}^{\circ}, Y_{i۲}^{m})$$

ساختارهای گم شدگی به صورت زیر می باشد[۳۹]

وقتی که گم شدگی به مقدارهای گمشده و مشاهده شده وابسته نباشد ، گم شدگی کاملا تصادفی [۷] (MCAR) است. در این حالت $P(R_{i۲}=۰|Y_{i۱}^{۰},Y_{i۲}^{۰})=P(R_{i۲}=۰)=p$ به عبارت دیگر احتمال این که پاسخ واحد iام در دور دوم که با نماد p نشان داده شده است، گمشده باشد به مقدار پاسخ مشاهده شده‌ی واحد iام در دور اول یا پاسخ واحد iام در دور دوم که گمشده است، بستگی ندارد.

در حالتی که گمشدگی متغیر پاسخ در دور دوم به دلیل پاسخی که در دور اول مشاهده شده رخ می دهد، گم شدگی تصادفی[۸] (MAR) است و $P(R_{i۲}=۰|Y_{i۱}^{۰},Y_{i۲}^{۰})=P(R_{i۲}=۰|Y_{i۱}^{۰})=p_i^{۰}$ یعنی احتمال گم شدگی پاسخ واحد iام در دور دوم وابسته به مقدار های مشاهده شده‌ی واحد iام در دور اول آمارگیری است که با نماد $p_i^{۰}$ احتمال گم شدگی پاسخ واحد iام در دور دوم که به داده های مشاهده شده وابسته است نشان می دهد.

اگر احتمال مقادیر گمشده، به مقادیر مشاهده نشده ای که باید به دست آید، وابسته باشد گم شدگی را غیر تصادفی[۹] (NMAR) است، که در آن p_i^m احتمال گم شدگی واحد iام که به داده های گمشده وابسته است را نشان می دهد.

$$P(R_{i۲}=۰|Y_{i۱}^{۰},Y_{i۲}^{۰})=p_i^m$$

۱-۱-۲ الگوریتم EM[۱۰]

راه های متعددی برای جانهی به داده های گمشده وجود دارد ازجمله روش جانهی میانگین ، روشی که به طور گسترده برای داده های گمشده مورد توجه قرار گرفته، جایگزین کردن میانگین به جای داده گمشده است، بدین صورت که میانگین هر متغیر به صورت جداگانه (بدون در نظر گرفتن داده

۱ -Missing Completrly at Random
۲- Missing at Random
۳- Not Missing at Random
٤- Expectation Maximization

های گمشده) محاسبه می گردد و آن‌گاه میانگین ها به جای داده های گمشده در همان متغیرها قرار داده می شوند. یا روش جانهی رگرسیون پیش بینی شده که در این روش با استفاده از داده های مشاهده شده رابطه رگرسیونی پاسخ بر حسب متغیرها برآورد می شود و سپس معادله رگرسیونی برای گمشدها به کار گرفته می شود تا پاسخهای آنها پیش بینی شوند. مدل های رگرسیونی به کار رفته برای انواع مختلف داده ها عبارتند از:

پیوسته رگرسیون خطی

شمارشی رگرسیون پواسن

روش دیگر جانهی گمشده ها الگوریتم EM است . الگوریتم EM یک روش معمول برای ماکسیمم کردن درستنمایی های پیچیده در برخورد با مسئله داده های شامل مشاهده های گمشده است. دامنه وسیعی از مسئله های آماری می تواند توسط این الگوریتم حل شود. [۳۹] از جمله کاربردهای این الگوریتم تعدیل اثر بی پاسخی می باشد.

دو فرض زیر برای استفاده از این الگوریتم در داده های گمشده مورد نیاز می باشند:

۱. پارامتر θ از پارامترهای فرایند گمشده مستقل است، و

۲. داده های گمشده دارای ساختار گم شدگی تصادفی باشند.

این الگوریتم بر اساس ایده ای بی قاعده برای برخورد با داده های ناتمام فرمول بندی شده است، به طوری که:

۱. مقدارهای گمشده را توسط مقدارهای برآورد شده جایگذاری می کند.

۲. پارامترها را برآورد می کند.

۳. مقدارهای گمشده را با برآوردهای جدید پارامترها دوباره برآورد می کند.

۴. مجدداً پارامترها را برآورد می کند و تکرار این مراحل تا همگرایی ادامه می یابد.

برآورد داده های گمشده در مرحله آخر به عنوان جانهی مقادیر گمشده استفاده می شود. برای استفاده از الگوریتم EM مجموعه داده ها باید به دو مجموعه داده تقسیم شود: اول مجموعه داده های مشاهده شده (داده های ناتمام) و دوم مجموعه داده های غیر قابل مشاهده (داده های کامل). در واقع مجموعه داده های ناتمام، مستقیماً مشاهده می شوند و مجموعه داده های کامل همراه با تعدادی داده‌ی گمشده می باشد. به بیان دیگر اگر y نماینده‌ی داده های ناتمام و x نماینده داده های کامل باشد، داریم $y = h(x)$ ، یعنی y کاملاً توسط x تعیین می شود، ولی عکس آن درست نیست.

فرض کنید که تابع درستنمایی داده های ناتمام و داده های کامل به ترتیب با $L(\theta, y)$ و $L(\theta, x)$ نشان داده شوند که θ پارامتر توزیع تحت بررسی است. برای معرفی الگوریتم EM ابتدا رابطه‌ی زیر تعریف می شود:

$$Q(\theta' \mid \theta) = E[\log f(X \mid \theta') \mid y, \theta]$$

که θ پارامتر در مرحله قبل و θ' پارامتر در مرحله جدید را نشان می دهد. فرض می شود که برای تمام زوج های (θ', θ) تابع Q وجود دارد. همچنین فرض می شود که $f(x \mid \theta) > ۰$ برای هر $\theta \in \Theta$ و x برقرار باشد. در این صورت الگوریتم EM در مرحله $k+۱$ ام به صورت زیر تعریف می شود:

گام E: مقدار $Q(\theta' \mid \theta^{(K)})$ محاسبه می شود.

گام M: مقدار $\theta^{(K+۱)}$ که $\theta \in \Theta$ طوری تعیین می شود که تابع $Q(\theta' \mid \theta^{(K)})$ را ماکسیمم کند.

نکته‌ی قابل توجه این است که ماکسیمم سازی تحت اولین شناسه‌ی تابع Q صورت می گیرد. در اینجا ایده‌ی اصلی به این صورت است که برای تعیین کردن برآورد ماکسیمم درستنمایی پارامتر θ نیاز به ماکسیمم کردن $\log f(x|\theta)$ می باشد، اما به دلیل این که $\log f(x|\theta)$ به طور کامل مشخص نیست لذا به جای آن امید ریاضی $\log f(x|\theta)$ به شرط داده های ناتمام $y, Q^{(k)}$ ماکسیمم می شود. در واقع گامE، این الگوریتم امید ریاضی شرطی داده های گمشده به شرط داده های مشاهده شده و پارامترهای برآورد شده‌ی جاری را محاسبه می‌کند و سپس این امید ریاضی را برای داده های گمشده جایگزین می‌کند، که همان جانهی داده های گمشده می باشد. گام M پس از این که مقدارهای گمشده جایگزین شدند، مانند به دست آوردن برآورد ماکسیمم درستنمایی به سادگی محاسبه می‌شود.

۱-۲ داده های دورافتاده

یک مشاهده دورافتاده داده‌ای است که به طور آشکار از دیگر اعضای داده‌ها فاصله دارد [۱۴] به بیان دیگر یک مشاهده دورافتاده مشاهده‌ای است که فاصله زیادی از مشاهدات دیگر داشته به طوری که گمان می‌رود که از مکانیسم متفاوتی تولید شده است.[۱۶]

علت لزوم در نظر داشتن دورافتاده‌ها چیست؟

تلاش برای شناسایی دورافتاده‌ها به دو دلیل عمده صورت می‌گیرد. دلیل اول این است که گاهی دور افتاده‌ها به جامعه دیگری تعلق دارند که باید شناسایی و حذف گردند. دلیل دوم این که ممکن است دورافتاده‌ها روی نتایج حاصل از بقیه داده‌ها اثرگذار باشند. برای مثال موجب اریبی برآوردها و تورم واریانس شوند. بنابراین به دنبال روش‌هایی هستیم که با کاهش اثرات داده های دورافتاده مشکلات مذکور به حداقل ممکن برسد. چنین روش‌هایی را روش‌های استوار[۱۱] می‌نامند. کلمه استوار به معنی حساس نبودن به انحراف اندکی از فرضیه‌ها به کار می‌رود.[۲۱]

۱-Robust

داده‌های دورافتاده چگونه پدید می‌آیند؟

دلایل متفاوتی موجب حضور داده‌های دورافتاده در نمونه می‌شود. برخی اوقات خطای اندازه‌گیری مانند خطا در ثبت داده‌ها و یا محدودیت ابزار اندازه‌گیری باعث به وجود آمدن دورافتاده‌ها می‌شود. مواقعی نیز وجود دارد که پراکندگی[۱۲] طبیعی موجود در یک مجموعه داده، دورافتاده‌ها را پدید می‌آورد. زمانی که تحلیل‌گر می‌خواهد برای حذف و یا عدم حذف یک مشاهده با مقدار فرین[۱۳] تصمیم بگیرد بایستی دلایل فوق را در به وجود آمدن آنها در نظر داشته باشد. اگر دورافتاده‌ها در اثر خطای اندازه‌گیری و یا خطای اجرای کار به وجود آمده باشد احتمالاً باید از نمونه حذف شوند. اما اگر پراکندگی موجود در خود داده‌ها، دورافتاده (یا دورافتاده‌هایی) را پدید آورده باشد، نمی‌توان آنها را حذف کرد.

چشم‌پوشی از داده‌های دورافتاده چه عواقبی به دنبال دارد؟

وجود یک یا چند دورافتاده باعث ایجاد اثر پوششی[۱۴] (بیرون‌بری) و اثر درون‌آوری[۱۵] می‌شود [۱۶]. بیرون‌آوری زمانی اتفاق می‌افتد که گروهی از دورافتاده ها برآوردگرهای میانگین و کوواریانس را به طرف خود می کشانند و در نتیجه فاصله نقاط دورافتاده از میا نگین کوچک خواهد شد. اثر درون‌آوری بدین صورت است که گروهی از مشاهدات، میانگین و کوواریانس را به سوی خود سمت داده و فاصله نقاط دیگر از میانگین بزرگ خوا هد شد. دو مشکل برون‌بری و درون‌آوری با استفاده از برآوردگرهای استوار مکان و مقیاس حل خواهد شد چرا که برآوردهای استوار تحت تأثیر دورافتاده‌ها نیستند. امروز محققان دریافته‌ا ند که پیش از به کارگیری روش‌های کلاسیک مانند میانگین نمونه و کوواریانس نمونه باید روش‌های شناسایی دورافتاده‌ها را استفاده کنند. از این رو به کارگیری روش‌های استوار که دورافتاده‌ها را شناسایی و لحاظ می‌کنند منطقی‌تر و کارآمدتر است.

۲-Variation
۳-Extreme
۴-Masking Effect
۵-Swamping Effect

داده‌های دورافتاده چندمتغیره

همچنان کـه مطالعـه دورافتـاده‌هـا در نمونـه‌هـای تـک‌متغیـره اهمیـت دارد، در مجموعـه داده‌هـای چندمتغیره نیز از اهمیت زیادی برخوردار است . شاید بتوان گفـت مشـاهدات دورافتـاده در مجموعـه داده‌های چندمتغیره از اهمیت بیشتری برخوردار است چرا که در چنین مواقعی دورافتاده‌ها خود را به راحتی نشان نمی‌دهند. هر چه بعد داده‌ها افـزایش یابـد ایـن مشـکل نیـز بیشـتر مـی‌شـود چـرا کـه دورافتاده‌ها در ابعاد مختلف پراکنده می‌شوند. در مسیر مطالعه دورافتاده‌های چندمتغیره، از معیاری به نام فاصله ماهالانوبیس[۱۶] می‌توان کمک گرفت.

۱-۳ فاصله ماهالانوبیس

نمونه تصادفی p -متغیره $X = \{(x_{i1},...,x_{ip})^t\}, i = 1,...,n$ را در نظر بگیرید. اگر μ_x میـانگین نمونـه متغیرهای X و Σ_{xx} ماتریس کوواریانس آن باشد، مربـع فاصـله آمـاری هـر مشـاهده x_i از μ_x بـه صورت

$$(x_i - \mu_x)'(\Sigma_{xx})^{-1}(x_i - \mu_x)$$

خواهد بود. اگر کلیه مشاهداتی را که مربع فاصله آنها از μ_x مقدار ثا بتی است به صورت

$$\{x_i : (x_i - \mu_x)'(\Sigma_{xx})^{-1}(x_i - \mu_x) = c^2\}$$

نمایش دهیم، مکان هندسی مشاهدات، بیضی‌واری با مرکز μ_x خواهد شد که بیضـی‌وار تحمـل[۱۷] نـام داشته و مشاهداتی را که خارج از بیضی‌وار قرار می‌گیرند به عنوان داده دورافتاده در نظر می‌گیریم.

بنابراین فاصله ماهالانوبیس هر مشاهده x_i را تعریف می‌کنیم:

$$MD(x_i) = \sqrt{(x_i - \mu_x)' (\Sigma_{xx})^{-1}(x_i - \mu_x)}$$

فاصلهٔ ماهالانوبیس به طور دقیق بر پایه آماره هایی بنا شده است که بیشترین حساسیت را نسـبت بـه داده‌های دورافتاده دارند. لذا به منظور دستیابی به معیارهای تشخیصی معتبرتر برای نقاط دورا فتـاده،

۱- Mahalanobis Distance
۱-Tolerance Ellipse

پیشنهاد می‌شود μ_x و Σ_{xx} را با برآوردهایی جایگزین کنیم که در حضور نقاط دورا فتاده مقاوم باشند. برآوردگرهای اخیر را که برآوردگرهای استوار نام دارند در بخش های بعدی معرفی می‌کنیم.

۱- ۴ برآوردگر استوار

بسیاری از روش‌های آماری براساس مفروضات ویژه ای نظیر نرمال بودن توزیع داده‌ها بنا شده است. اما اگر داده ها از توزیع نرمال پیروی نکنند با مشکل مواجه هستیم. به بیان دیگر بیشترین مشکل زمانی رخ می‌دهد که بعضی از مشاهدات، دورافتاده یا گم شده باشندروش های کلاسیک مانند کمترین مربعات نسبت به این نقاط بسیار حساس می باشند. بنابراین هدف اصلی، استفاده از روش‌های استوار به علت استواری آنها درحضور نقاط دورا فتاده وگم شده است. برآوردهای چندمتغیره مکان و مقیاس، از مهمترین مسائل در آماره های استوار هستند [۳۶-۳۵-۲۱-۱۱]. روش‌های استوار قادرهستند به مشاهدات وزن‌های نابرا بر اختصاص دهند. به طور کلی مشاهداتی که مانده‌های بزرگی را تولید کنند به وسیله برآوردگر استوار وزن کم‌تر داده می‌شوند.

۱-۵ خواص مناسب برای یک برآوردگر استوار

۱- ۵ -۱ نقطه فروریزش(BDP)[۱۸]

برخی برآوردگرها نسبت به نقاط دور افتاده حساس بوده وگاهی تنها یک نقطه دور افتاده برای مغشوش کردن آنها کافی است. میزان عدم حساسیت یک برآوردگر نسبت به نقاط دورافتاده، نقطه فروریزش نامیده می شود.[۱۰] ، نقطه فروریزش نمونه متناهی را به صورت زیر تعریف کرده ا ند:

۱-Breakdown point

مجموعه داده های $Z_n = (X,Y) \in R^{n\times(p+q)}$ را در نظر بگیرید. اگر T_n برآوردگر و $T_n(Z_n')$ برآوردگر تمام نمونه های تحریف شده Z_n' باشد که به وسیله m مقدار دلخواه به جای مقادیر اصلی بدست آمده است، بیشترین اریبی که با ورود یک ناخالصی ایجاد می گردد برابر است با:

$$\sup_{Z_n'} \|T_n(Z_n) - T(Z_n')\| = bias(m, T_n, Z_n)$$

اگر این مقدار نامتنا هی باشد، به این معنا که m داده دورافتاده می توا ند روی برآوردگر T_n تأثیر زیادی داشته باشد لذا برآوردگر را مغشوش می کند. نقطه فروریزش برآوردگر T_n در نمونه متناهی Z_n' به صورت

$$\varepsilon_n^*(T_n, Z_n) = \min_{1\le m\le n}\left\{\frac{m}{n} : \sup_{Z_n'} \|T_n(Z_n) - T(Z_n')\| = \infty\right\}$$

تعریف می شود. به عبارت دیگر نقطه فروریزش، کوچکترین کسری از ناخالصی هاست که سبب می‌شود برآوردگر T_n مقداری دورتر از مقدار واقعی $T_n(Z_n)$ بگیرد. به طور کلی هرچه نقطه فروریزش یک برآوردگر کمتر باشد، میزان حساسیت آن نسبت به داده های دورافتاده بیشتر خواهد بود و بالعکس. بنابراین به منظور کا هش میزان حساسیت نسبت به نقاط دورافتاده، به دنبال برآورد گرهایی با نقاط فروریزش بالا هستیم.

می توان نشان داد که نقطه فروریزش میانگین نمونه برابر صفر و نقطه فروریزش میانه ۰/۵ است. [۱۰]

اکنون با ارائه تعاریف "بردارهای مستقل خطی"، "مقادیر و بردارهای ویژه یک ماتریس" و "نامنفرد بودن ماتریس مربع A" ، نقطه فرو ریزش برآوردگر ماتریس کوواریا نس را تعریف خوا هیم کرد.

تعریف ۱-۱:

مجموعه بردارهای $X_1, X_2, ..., X_k$ را وابسته خطی گویند هرگاه k ثابت، $c_1, c_2, ..., c_k$ وجود داشته باشد که حداقل دوتا ازآنها غیر صفر باشد به طوری که

$$c_1X_1 + c_2X_2 + ... + c_kX_k = 0 \qquad (1-1)$$

درغیر این صورت مجموعه بردارها را مستقل خطی نامند.

تعریف ۱-۲:

ماتریس مربع $A_{k\times k}$ دارای مقدار ویژه λ وبرداروویژه $x_{k\times 1} \neq 0$ متناظربا λ است هرگاه

$$Ax = \lambda x \qquad x \neq 0 \qquad (2-1)$$

برای بدست آوردن مقادیر ویژه A به صورت زیر عمل می کنیم:

$$Ax = \lambda x \quad \Leftrightarrow Ax - \lambda x = 0 \quad \Leftrightarrow (A - \lambda \mathrm{I})x = 0$$

$\Leftrightarrow A - \lambda \mathrm{I}$ وارون پذیر نباشد

$\Leftrightarrow \quad |A - \lambda \mathrm{I}| = 0$

بنا براین باحل معادله $|A - \lambda \mathrm{I}| = 0$، k مقادیر ویژه A به صورت $\lambda_1, \lambda_2 ..., \lambda_k$ به دست می آیند که با قرار دادن این مقادیر ویژه در دستگاه $Ax = \lambda x$، بردارهای ویژه ماتریس نیز بدست می آیند .معمولاً بردارهای ویژه بعد از محاسبه نرمال می شوند به گونه ای که طول آنها یک شود. این کار به دلیل سهولت در محاسبات و ایجاد شرط استقلال خطی بین بردارهای ویژه ا نجام می پذیرد. بردارهای نرمال شده را با $e_1, e_2, ..., e_k$ نشان می دهیم.

تعریف ۱-۳

یک ماتریس مربع $A_{k\times k}$ نامنفرد(وارون پذیر)است هرگاه دستگاه $A_{k\times k} x_{k\times 1} = 0$ تنها دارای جواب بدیهی $x_{k\times 1} = 0$ باشد. به عبارت دیگر ماتریس مربع نامنفرد است هر گاه ستون ها (سطرهای) آن مستقل خطی باشند.

حال مقدار فروریزش برآوردگر ماتریس کوواریانس $C(Z_n)$ راکوچکترین کسر از ناخالصی ها است تعریف می کنیم که بزرگترین مقدار ویژه $C(Z_n')$ را به طوردلخواه بزرگترین یا کوچکترین مقدار ویژه $C(Z_n')$ را به طوردلخواه کوچک(نزدیک صفر)کند.یعنی:

$$\varepsilon_n^*(C_n, Z_n) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup D(C_n(Z_n), C_n(Z_n')) = \infty \right\} \qquad (۳-۱)$$

که در آن

$$D(A,B) = \max\{|\lambda_1(A) - \lambda_1(B)|, |\lambda_p(A)^{-1} - \lambda_p(B)^{-1}|\} \qquad (۴-۱)$$

که در آن $\lambda_1(A) \ge \lambda_2(A) \ge \ldots \ge \lambda_p(A)$ مقادیر ویژه ماتریس A هستند.

۱-۵-۲ هم وردایی افاین[۱۹]

یک شرط ضروری برای برآوردگر چند متغیره هم وردایی تحت تبدیلات افاین است. فرض کنید $X_{n,p} = \{X_1, X_2, \ldots, X_n\} = \{(x_{11}, x_{12}, \ldots, x_{1p}), \ldots, (x_{n1}, x_{n2}, \ldots, x_{np})\}$ یک مجموعه از n نقطه در p بعد باشد و $T(X) \in R^p$ بردار مکان براساس X است. می گوییم $T(X)$ هم وردای تبدیل است هر گاه

$$T(X + b) = T(X) + b$$

برای هر بردار p بعدی b که $X + b = \{X_1 + b, \ldots, X_n + b\}$.

هم وردایی افاین ترکیبی از تبدیل خطی و انتقال است یعنی $T(XA+b) = T(X)A + b$

که در آن A هر ماتریس منفرد $p \times p$ و b هر بردار p بعدی است. به تعبیر هندسی هم وردایی افاین نگاشت خطوط مستقیم به خطوط مستقیم می باشد، یا به عبارت دیگر خط بعد از تبدیل خط باقی می ماند.

یک برآورد کوواریا نس $C(X)$ هم وردای افاین است هرگاه

$$C(XA+b) = A'C(X)A$$

ویا به عبارت دیگر

$$C(\{X_1A + b, \ldots, X_nA + b\}) = A'C(X_1, \ldots, X_n)A \qquad (۶-۱)$$

میا نگین نمونه وماتریس کوواریانس نمونه هردو هم وردای افاین هستند زیرا داریم:

۱-Affine equivariant